Genome-to-genome analysis reveals the impact of the human innate and adaptive immune systems on the hepatitis C virus

M. Azim Ansari[1,2#], Vincent Pedergnana[1#], Camilla Ip[1], Andrea Magri[2], Annette Von Delft[3], David Bonsall[3], Nimisha Chaturvedi[4], Istvan Bartha[4], David Smith[3], George Nicholson[5], Gilean McVean[1,6], Amy Trebes[1], Paolo Piazza[1], Jacques Fellay[4], Graham Cooke[7], Graham R Foster[8], STOP-HCV Consortium, Emma Hudson[3], John McLauchlan[9], Peter Simmonds[3], Rory Bowden[1], Paul Klenerman[3], Eleanor Barnes[3] & Chris C. A. Spencer[1*]

[1] Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford, OX3 7BN, UK

[2] Oxford Martin School, University of Oxford, 34 Broad Street, Oxford, OX1 3BD, UK

[3] Nuffield Department of Medicine and the Oxford NHIR BRC, University of Oxford, Oxford, OX1 3SY, UK

[4] School of Life Sciences, École Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland

[5] Department of Statistics, University of oxford, Oxford, OX1 3LB, UK

[6] Oxford Big Data Institute, Li Ka Shing Centre for Health Information and Discovery, University of Oxford, OX3 7BN, UK

[7] Wright-Fleming Institute, Imperial College, London, UK

[8] Queen Mary University of London, 4 Newark Street, London, E1 4AT, UK

[9] Centre for Virus Research, Sir Michael Stoker Building, 464, Bearsden Road, Glasgow, G61 1QH, UK

# Equal contribution

* Corresponding author

Abstract

Outcomes of hepatitis C virus (HCV) infection and treatment depend on viral and host genetic factors. We use human genome-wide genotyping arrays and new whole-genome HCV viral sequencing technologies to perform a systematic genome-to-genome study of 542 individuals chronically infected with HCV, predominately genotype 3. We show that both HLA alleles and interferon lambda innate immune system genes drive viral genome polymorphism, and that *IFNL4* genotypes determine HCV viral load through a mechanism that is dependent on a specific polymorphism in the HCV polyprotein. We highlight the interplay between innate immune responses and the viral genome in HCV control.

Introduction

Hepatitis C virus (HCV) infection presents a major health burden, infecting more than 185 million people worldwide[1] and leading to liver failure and hepatocellular cancer. Both host and virus genetic variations are associated with important clinical outcomes. Host genetic polymorphisms, most notably in the interferon lambda 3 and 4 locus, are associated with spontaneous clearance of the virus, response to treatment, viral load and progression of liver disease[2–6]. Similarly, viral genotypes, and distinct viral genetic motifs have been associated with the response to interferon based therapies[7,8] whilst resistance-associated substitutions (RASs) have been identified for most of the new oral direct-acting antiviral (DAA) drugs[9–12]. HCV can be divided into seven major genotypes, and most of the genetic data acquired to date focuses on HCV genotype 1 with a paucity of data addressing other genotypes. HCV genotype 3 is of particular interest as this genotype is known to infect 53 million people globally[13], has particularly high prevalence in the UK, and is associated with a higher failure rate to DAA therapies[14,15]. Despite these key observations, there has been few previous attempts to generate large full-length HCV genome datasets and to broadly assess the impact of host genes on the HCV viral genome.

Previous work has shown that within-host virus diversity evolves in response to the adaptive immune system, including candidate genes studies of the association between the human leukocyte antigen (HLA) type I proteins with the HCV genome[16,17]. HLA molecules are expressed on most cell types and present viral peptides (epitopes) to cytotoxic T lymphocytes (CTLs) that kill infected cells. CTL-mediated killing of virus-infected cells drives the selection of viral polymorphisms

("escape" mutants) that abrogate T cell recognition[18]. Understanding how host HLA molecules impact on viral selection has important implications for the development of HCV T cell vaccines that aim to prevent infection[19,20]. Since both HLA molecules and the viral epitopes they present are highly variable, a comprehensive host genome to viral genome analysis at scale is needed to assess the relative contribution of host HLA molecules in driving HCV viral genetic change. An analysis of this kind may also reveal other host genes that play a key role in shaping the HCV viral genome.

Whole genome viral sequencing for HCV is particularly challenging since HCV is genetically highly diverse both within and between infected individuals[21,22]. To overcome this we have recently developed new technologies that use next generation sequencing with targeted enrichment[23,24] to obtain a consensus genome in each individual at scale. The falling cost of host genotyping, in combination with large reference panels of human genomes[25] and advances in statistical methodologies, mean that it is possible to obtain accurate inference of millions of host genetic polymorphisms, including classical HLA alleles[26–28]. For the first time, these developments allow an unbiased genome-to-genome analysis[29], in large cohorts of HCV infected individuals, to better understand how host genetic variation drives changes within the virus during persistent infection at a population level.

We generated data from a cohort of 601 HCV infected patients (recruited to a clinical study called BOSON[30]) to systematically look for associations between host and virus genomes, exploiting the fact that while host genetics remain fixed, the virus mutates, allowing it to evolve during infection. We provide evidence that polymorphisms relevant to the innate (*IFNL4*) and adaptive immune systems (HLA genes) are associated with HCV sequence polymorphisms (so-called "footprints" in the viral genome). We show that an interaction between host *IFNL4* genotype determines HCV viral load through a mechanism that is dependent on a specific polymorphism in the HCV polyprotein. By assessing rates of viral evolution in individuals with different *IFNL4* genotypes, we highlight systematic differences in the innate immune response and discuss how these might relate to previous associations with spontaneous clearance and clinical treatment. We discuss the potential for a joint analysis of host and viral genome data to provide information on underlying molecular interactions and their importance in treating and preventing HCV, and other viral infections, in the era of genomic analysis.

Results

Sample description and genetic structure

DNA samples from 567 patients within the BOSON clinical study[30] (out of 601 patients) were genotyped using the Affymetrix UK Biobank array. This array genotypes over 800,000 bi-allelic single nucleotide polymorphisms (SNPs) across the human genome, including a set of markers specifically chosen to capture common HLA alleles. Pre-treatment plasma samples from 583 patients in the same study were analysed to obtain HCV whole genome consensus sequences using a high-throughput HCV targeted sequence capture approach coupled with Illumina sequencing[23].

Both full-length HCV genome sequences and human genome-wide SNP data were obtained on a total of 542 patients of mainly White and Asian self-reported ancestry infected with HCV genotypes 2 or 3 (see **Supplementary Table S1** for patient demographics). After quality control and filtering of the human genotype data, approximately 330,000 common SNPs with minor allele frequency greater than 5% were available for analysis along with inferred alleles at both class I and II HLA genes. The full-length HCV genome is approximately 9.5 kb, corresponding to over 3000 encoded amino acids. In our dataset, 1226 sites of the HCV proteome were defined to be variable (where at least 10 isolates have an amino acid which differs from the consensus amino acid) to have minimal statistical power for analysis.

We characterised human genetic diversity in the cohort via principal component analysis (PCA). The first two principal components (PCs) corresponded to the sample's 83% White and 14% Asian self-reported ethnicity, (**Supplementary Figure S1**), which differed significantly in some of the inferred HLA alleles frequencies (**Supplementary Table S2**) consistent with previous observations[31]. The third PC separates individuals with Black self-reported ethnicity from the rest of the cohort. We summarised virus diversity by constructing a maximum-likelihood tree of the consensus sequences from each patient (**Supplementary Figure S2**). Major clades in the tree separated HCV genotypes 2 (8% of the sample) and 3 (which in turn comprised clades representing subtype 3a and non-subtype 3a samples (90% and 2% of the total respectively)). A PCA on virus nucleotides sequence data reflected the structure of the tree, specifically at the level of virus subtypes (**Supplementary**

Figure S3). We used the PCs as covariates to control for viral population structure in the primary genome-to-genome analysis.

Labelling the inferred tree of viral diversity with either genetic ancestry (measured by host PCs) or self-reported ethnicity revealed a group within genotype 3 that was strongly associated with Asian ancestry (**Supplementary Figure S2**). The viral group associated with hosts of Asian ancestry is basal on the phylogenetic tree relative to the viral clade associated with White ancestry. The observed relative reduction in viral diversity within hosts of White ancestry is likely to be explained by the introduction and spread of HCV genotype 3a from South Asia where it is endemic[13].

Systematic host genome to virus genome analysis

We used the genotyped autosomal SNPs in the host genome to undertake genome-wide association studies, where the traits of interest were the presence or absence of each amino acid at the variable sites of the virus proteome, resulting in nearly one billion association tests. The analyses were performed on a compute cluster using logistic regression as implemented in PLINK2[32], adjusting for sex, population structure, and assuming an additive model. To control for population structure, we included the first three PCs of the host and the first 10 PCs of the virus as covariates. We note that failure to control for either covariates leads to a significant inflation in the association test statistics (**Supplementary Figure S4**), as would be expected given the observed correlation in population structure of the virus and the host (**Supplementary Figures S1, S2 and S3**). Assuming a human genome-wide significance threshold of 5×10^{-8} [33], and that amino acid variants in the viral genome are approximately uncorrelated once the population structure is accounted for, then a Bonferroni correction[34] results in a significance threshold of approximately 2×10^{-11}.

Figure 1 shows the result of the genome-to-genome association tests and highlights the strongest signals of association. Across the human genome, the most significant associations were observed between multiple SNPs in the major histocompatibility complex (MHC; chromosome 6) locus and a virus amino acid variant in non-structural protein 3 (NS3). Three other associations were observed between multiple SNPs in the host MHC and virus amino acids in NS3 and NS4B proteins ($P<2 \times 10^{-9}$). Outside the MHC, the strongest association between host and virus was detected between the SNP rs12979860 in the *IFNL4* gene (chromosome 19) and HCV amino acids at position 2570 in NS5B protein ($P=1.98 \times 10^{-9}$). Observed variability in the

density of nominally significant associations (**Figure 1**) is largely explained by variability in host and virus sequences, for example in the hyper-variable region (HVR) of HCV in E2 protein.

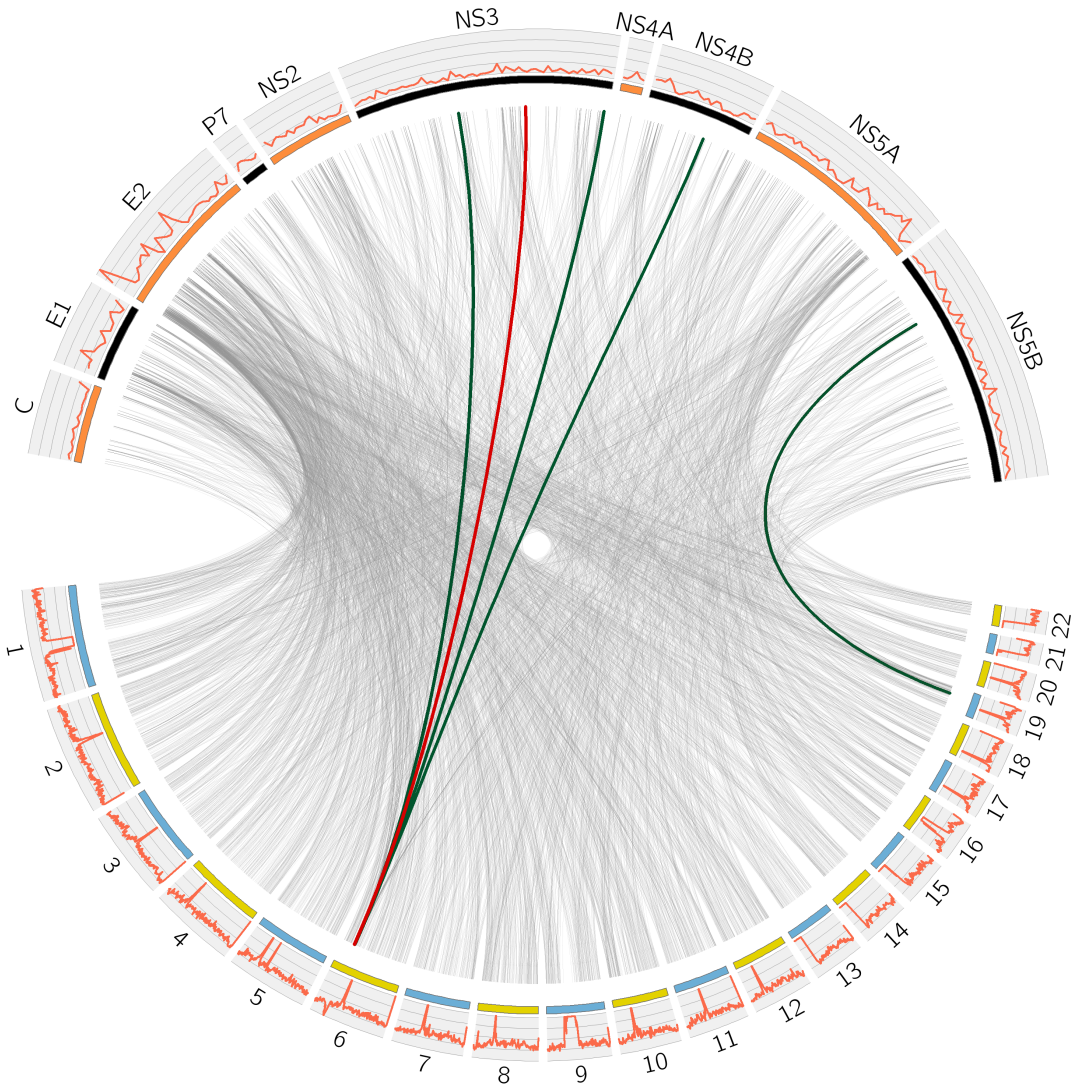


Figure 1. Human to hepatitis C virus genome-wide association study in 542 patients. The lower arc shows the human autosomes from chromosome 1 to 22, and the upper arc shows the HCV proteome from core (C) to NS5B. The red line represents the most significant association, $P<2x10^{-11}$. The four green lines represent suggestive associations, $P<2x10^{-9}$. The thin grey lines represent associations with $P<10^{-5}$. The outer mini-panels represent, on the upper arc, the viral diversity as measured by Shannon entropy and, on the lower arc, the density of human SNPs in bins of 1Mb, with higher values away from the centre for both upper and lower arcs.

We observed 169 associations between human SNPs mapping to two loci and five HCV amino acid sites ($P<4x10^{-9}$) (**Supplementary Table S3**). Since these associations represent places where host genetic diversity has impacted on virus sequence diversity we refer to them as “footprints”. We interpret the signals of

association in the MHC region to indicate the effect of the adaptive immune system on genetic diversity in the virus genome. Whilst the effect of MHC was anticipated, the strong signal of association of the interferon lambda region with viral diversity indicates additional effects of the innate immune response.

HLA alleles to virus genome

SNPs in the MHC which show strong association with viral amino acids are likely to be correlated with alleles at HLA genes due to extensive linkage disequilibrium across the region[35,36]. The HLA repertoire of a patient defines which viral peptides will be presented to T cells as part of the adaptive immune response, and can lead to the selection of viral mutations away from these peptides ("escape mutations")[16,37–40]. Upon transmission to another host, with a different HLA repertoire, reversion to the wild type may occur ("reversion mutations").

Using the tree inferred from whole-genome viral sequences we estimated the ancestral amino acids at internal nodes of the tree[41]. Inferring the changes along the terminal branches of the tree aims to control for the confounding between host and virus population structures[42] by looking at viral mutations after infection. To test for association between specific HLA alleles and viral amino acids, we constructed a 2x4 contingency table for each amino acid present at variable sites, and counted the number of changes and non-changes from ancestral amino acids in carriers, and non-carriers, for each imputed HLA allele (see **Figure 3a** for an example). We used a Fisher's exact test to assess significance of the association along the viral genome (**Figure 2a**), and a permutation approach to estimate the false discovery rate (FDR) as the *P*-values for discrete counts with low number of observation was not uniform in distribution[43,44].

At a 5% FDR, 24 combinations of HLA alleles and HCV sites were significant (**Table 1**) and this increased to 153 associations at a 20% FDR (**Supplementary Table S4**). Out of 21 viral amino acid positions showing signals of association with one or more HLA alleles, 12 were located in previously reported HCV genotype 3 epitopes[20] **(Table 1**), which represents a strong enrichment (odds ratio, OR=5.2, P=2.8x10^{-4}). We also observed that the NS3 protein was strongly enriched for association signals with HLA alleles (OR=6.6, P=5.68x10^{-5}) and that three HLA alleles were nominally enriched for association signals with viral amino acids (*HLA-A*31:01*, OR=10.5, P=0.02; *HLA-A*32:01*, OR=13, P=0.01; *HLA-A*68:02*, OR=18.5, P=8x10^{-4}). The

strongest HLA footprinting signals are found with common alleles at *HLA-A* and *HLA-B* genes, although signals are also found in *HLA-C* and the class II gene *HLA-DQA1*. At position 1646 in the HCV genome (in NS3), the footprint is seen with multiple HLA alleles (*B*08:01*, C*07:01 and *DQA1*05:01*), although this is potentially a result of linkage disequilibrium between these HLA alleles (r^2 between *B*08:01* and *C*07:01* = 0.49, and r^2 between *B*08:01* and *DQA1*05:01* = 0.22; **Supplementary Figure S5**). As the majority of the cohort are patients infected with genotype 3a virus and self-reporting as White (411 out of 542), we repeated the HLA analysis on specifically these patients as an additional check that the associations are not driven by correlations between viral genotype and host ancestry. Broadly the association evidence was consistent with analysis of the whole cohort with some loss of power due to reduced sample size (**Supplementary Table S5** and **Supplementary Figures S6**) and potentially differential effects between the viral subtypes.

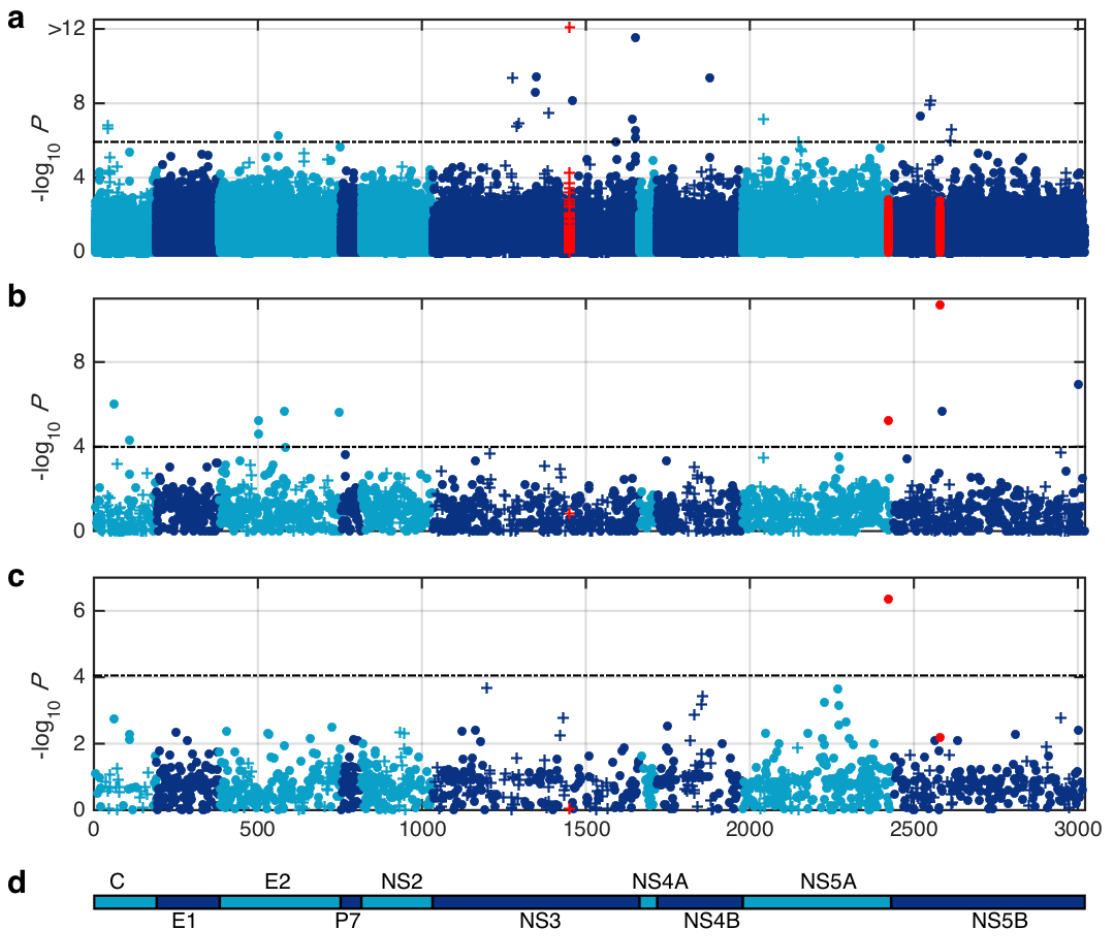


Figure 2. Association between HCV amino acids and (**a**) HLA alleles, (**b**) *IFNL4* genotypes using Fisher's exact test, and (**c**) pre-treatment viral load (log_{10}PTVL) using linear regression. Sites in experimentally validated epitopes in HCV genotype

3^{20} are indicated by a plus sign. Viral sites 1444, 2414 and 2570 are coloured as red. Dashed lines represent a 5% false discovery rate. (**d**) HCV polyprotein.

The most significant association was found for *HLA-A*01:01* and the viral amino acid at position 1444 in the NS3 protein. Having identified a viral amino acid subject to footprint by a host HLA allele, we can look for evidence of escape and/or reversion mutations at this viral site (**Figure 3a**). The *HLA-A*01:01* allele was associated with a tyrosine-to-phenylalanine change in the NS3 protein (Y1444F). Carriers of *HLA-A*01:01* allele very rarely lose an existing F amino acid, and there is a strong relative increase in the inferred number of changes from Y to F in individuals that carry *HLA-A*01:01* allele (P=1x10^{-32}), consistent with known T cell selection at this epitope[45]. The signal of association is at the ninth amino acid of the experimentally validated epitope (**Supplementary Figure S7**) which is known to be prone to escape mutations as it alters the contact between the epitope and the HLA groove which presents the antigen[46].

a

	Y -> Y	Y -> F	F -> F	F -> Y
A*01:01	4	13	151	0
non-A*01:01	105	12	172	84

b

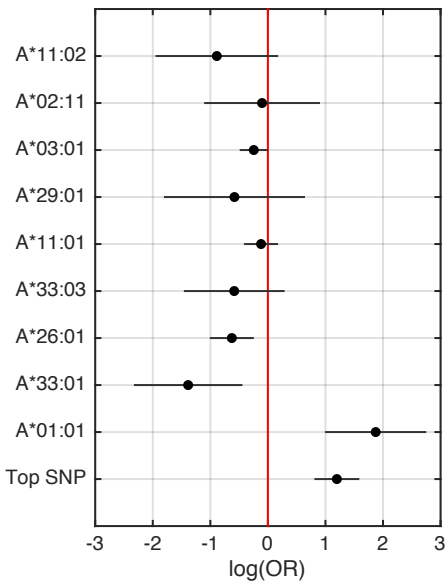


Figure 3. Association between viral position 1444 and *HLA-A* alleles. (**a**) Table of changes to and from the ancestral amino acids in individuals with and without the *HLA-A*01:01* allele. (**b**) Estimated effect size (log odds ratio) of common *HLA-A* alleles and the top associated SNP (rs3129073) on the possibility of mutation from F to Y at viral position 1444. The black dots represent the effect size estimate and the bar represent the 95% confidence intervals.

To quantify the influence of the host's HLA at position 1444, we estimated the odds ratio separately for the most common *HLA-A* alleles, and the most associated SNP in

the region (rs3129073), on the amino acid changes from F to Y. **Figure 3b** shows that our data support a very strong signal for individuals carrying *HLA-A*01:01* allele to carry a virus with the non-antigenic F amino acid which is consistent with the size of effect of the rs3129073. A weaker signal for *HLA-A*33:01* allele in the opposite direction is likely explained by these alleles acting as surrogates for non-*HLA-A*01:01* allele carriers. Association analysis across the HLA region which was conditioned on *HLA-A*01:01* allele (**Supplementary Figure S8**) removed all association signals ($P>10^{-4}$), suggesting that the primary *HLA-A*01:01* association explains other associations in the region with amino acids at position 1444.

Using the 27 HLA and viral amino acid associations that surpass the 20% FDR threshold, and have sufficient observations to estimate the ORs, we observed a negative correlation between the ORs of escape and reversion (R=-0.65, $P<0.01$; **Supplementary Figure S9**). These observations are consistent with HLA alleles driving patterns of both escape and reversion at viral amino acids. Our analysis provides a map of their influence across the HCV genome.

IFNL4 variants to virus genome

Variants in the interferon lambda region have been associated with multiple HCV outcomes including spontaneous clearance, treatment response, viral load and liver disease progression[2,3,47]. In our genome-wide analysis, variants around the interferon lambda region showed the strongest association with HCV amino acids outside the MHC region (the top associated SNP is rs12979860, $P=1.98\text{x}10^{-9}$). For that SNP, the CC genotype is associated with higher rates of spontaneous clearance and interferon-based treatment response, putatively due to the fact that it tags a deletion polymorphism (rs368234815) which prevents *IFNL4* expression[48]. In our cohort, only two individuals had discordant genotypes for rs12979860 and rs368234815 (which was imputed), resulting in a strong linkage disequilibrium ($r^2= 0.992$) between these two SNPs. Non-CC genotypes express *IFNL4* and have increased expression of interferon stimulated genes (ISGs)[49].

To interrogate these associations further, we compared viral amino acid changes between hosts with CC and non-CC genotypes at the *IFNL4* SNP rs12979860 using the same Fisher's exact test as described above. The most significant association was with changes to and away from valine (V) at position 2570 in the viral NS5B protein ($P=1.94\text{x}10^{-11}$) (**Figure 2b**). We replicated this association using an

equivalent analysis in an independent study of 360 patients of European ancestry chronically infected with HCV genotypes 2 or 3[4] (one-sided *P*=0.005). In addition, a candidate gene association study in a genotype 1b single source infection cohort has shown an association between *IFNL4* genotypes (rs12979860) and an amino acid in the same region of NS5B (position 2609)[38] reinforcing the potential role of the locus in interactions with the host immune system.

Overall, using a 5% FDR, 11 significant associations were observed between *IFNL4* genotypes and amino acid positions located in core, E2, NS5A and NS5B proteins (**Supplementary Table S6**). These amino acids are not located in any known HCV genotype 3 HLA-associated epitopes, however one of the sites (position 109 in the core protein) is also associated with *HLA-B*41:02* in our dataset (*P*=4.3x10^{-6}, **Table S4**). The top signals of associations between *IFNL4* genotypes and viral variants are located in viral proteins known to be HCV immune response targets[38], and those which disrupt human proteins in order to escape mechanisms of innate immunity[50]. We used a permutation approach to determine if any of the viral proteins are enriched in the number of associations with *IFNL4* genotypes. Only the core protein is nominally (*P*<0.05) enriched, while the NS5A protein is nominally depleted in the number of associations with the *IFNL4* genotypes.

In the 487 genotype 3a infected patients, we defined the population consensus to be the most common amino acid and assessed whether regions of the viral genome with either high diversity (HVR1, HVR2 and HVR3) or previously reported to be correlated with interferon based therapy outcome (the IFN sensitivity-determining region (ISDR), protein kinase phosphorylation homology domain (PePHD) and the protein kinase binding domain (PKR-BD))[7,51,52] showed difference in the number of changes from the population consensus, between patients with the CC and non-CC genotypes (**Supplementary Table S7**). Patients with the CC genotype had an increased number of mutations away from the population consensus in the ISDR (*P*=3.5x10^{-4}), HVR2 (*P*=2.5x10^{-6}) and PKR-BD (*P*=0.012) regions. However, we observed the same pattern across the genome (*P*=5.06x10^{-6}) suggesting that the *IFNL4* effect is not necessarily specific to these regions.

To further investigate the selective pressures on the virus in *IFNL4* CC and non-CC patients, we estimated the rates of synonymous (dS) and non-synonymous substitutions (dN) in genotype 3a infected patients (**Figure 4**). Whilst there was no difference in dS in CC and non-CC patients (**Figure 4a**, *P*=0.68), dN was significantly

higher in CC patients (**Figure 4b**, P=1.6 x10^{-8}). The lower dN/dS ratio in non-CC patients (P=1.3x10^{-10}) is potentially indicative that the virus is under a stronger purifying selection (**Figure 4c**). This hypothesis is supported by the observation that for the same rate of synonymous substitutions dS (a surrogate for the time and amount of divergence), the HCV genome is under a larger purifying selection in patients with *IFNL4* non-CC genotype than CC genotype (**Figure 4d**).

Estimating dN/dS ratio per viral gene showed that this ratio was significantly higher (P<0.05) in CC patients compared to non-CC patients in E1, E2, NS3 and NS5B (**Supplementary Figure S10**). A sliding window analysis across the HCV genome showed that in the full cohort E1 and E2 genes had a much higher dN/dS ratio compared to the rest of the genome. These envelope genes include hyper-variable regions (HVRs), and are thought to be the primary targets of the antibody-based immune response (**Supplementary Figure S11**).

We also tested whether *IFNL4* genotypes had an impact on HLA presentation as measured by our foot printing analysis. We did not find any consistent evidence for interaction between individual sites presentation by HLA alleles and *IFNL4* genotypes (**Supplementary Figure S12**), or in the mean number of escape mutations in CC and non-CC patients when comparing HLA allele carriers and non-HLA allele carriers.

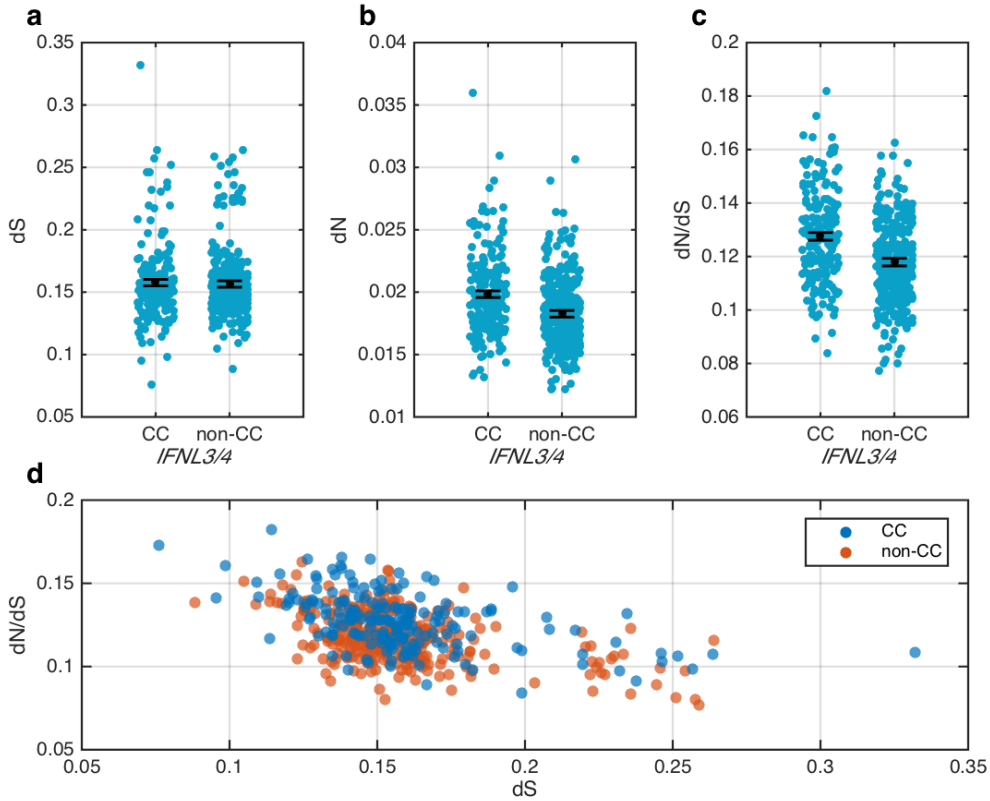


Figure 4. Association between *IFNL4* genotypes and rates of synonymous (dS) and non-synonymous (dN) substitutions in HCV genome in genotype 3a infected patients. Each blue dot represents the mean dS, dN or dN/dS ratio per patient. The mean and 95% confidence intervals are shown as black dots and bars. (**a**) Rate of synonymous substitution (P=0.68). (**b**) Rate of non-synonymous substitution (P=1.6x10^{-8}). (**c**) dN/dS ratio (P=1.28x10^{-10}). (**d**) The joint distribution of dS and dN/dS in individuals with the *IFNL4* non-CC genotype (red dots) and with the CC genotype (blue dots).

Host and virus genetic determinants of viral load

Pre-treatment viral load (PTVL, as measured in IU/ml) was available for all patients (See **Supplementary Table S1** for details of the cohort). We performed a genome-wide association study in patients infected with HCV genotype 3a using an additive linear regression model adjusted for sex and the three first host PCs for $\log_{10}$ transformed PTVL ($\log_{10}$PTVL) (**Figure 5a**). We replicated the known association between *IFNL4* variants on chromosome 19 and viral load[3] (rs12979860, P=5.9x10^{-10}) with the non-CC genotype conferring an approximately 0.45-fold decrease in viral load (mean for non-CC=3.4760x10^{6} IU/mL and for CC=6.3447x10^{6} IU/mL).

We also performed a genome-wide association study to detect associations between viral amino acids and viral load (**Figure 2c**). The only amino acid significantly associated with $\log_{10}$PTVL at a 5% FDR was a change from a serine (S) to an

asparagine (N) at position 2414 in NS5A protein (P=9.21x10^{-7}) (**Figure 5b**). This site is one of the 11 sites significantly associated (5% FDR) with *IFNL4* genotypes (**Figures 2b and 2c**). In patients with a serine at position 2414, the association between *IFNL4* genotypes and PTVL is highly significant (P=9.37x10^{-9}). However, we observed no association between (P=0.9) $\log_{10}$PTVL and *IFNL4* genotypes in patients infected with a virus that has a different amino acid (**Figures 5c and 5d**). In other words, the host's *IFNL4* genotypes determine viral load only if they are infected by a virus with the serine amino acid at position 2414 in NS5A protein (**Figures 5c and 5d**). The interaction is statistically significant when analysing either the whole cohort (P=0.0173) or just patients with genotype 3a infections who self-report as being White (P=0.01692). Together the combinations of a non-CC genotype and a serine viral amino acid at position 2414 are inferred to result in a 0.57-fold decrease in viral load compared to all other combinations (mean viral load for non-CC and serine at position 2414 =2.8083x10^{6} IU/mL and all other combinations =6.4735x10^{6} IU/mL).

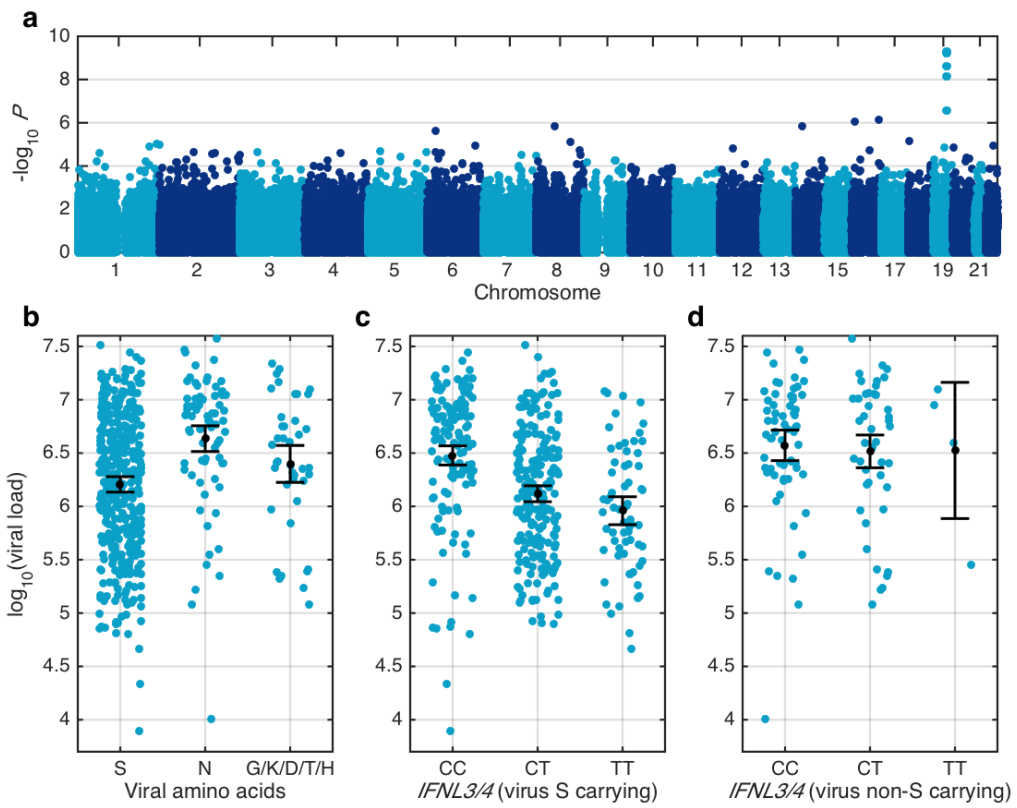


Figure 5. Association between $\log_{10}$ transformed pre-treatment viral load ($\log_{10}$PTVL) and human and virus genetic variants in genotype 3a infected patients. (**a**) Association between human SNPs and $\log_{10}$PTVL. (**b**) Distribution (blue dots), estimated mean and 95% confidence interval (shown as black dots and bars) of $\log_{10}$PTVL stratified by amino acids present at viral position 2414 (P=9.21x10^{-7}), (**c**) by *IFNL4* (CC, CT and TT) genotypes in patients whose virus carries a serine at

position 2414 (P=9.37x10^{-9}) (**d**) or in patients whose virus does not carry a serine amino acid at position 2414 (P=0.9).

To investigate whether the mutations associated with higher viral loads and/or *IFNL4* genotypes influence replication ability, each was introduced into the modified S52 replicon of genotype 3a[53] and replication assessed in short term Huh7.5 cell cultures. As a positive control, introduction of an alanine to threonine change at position 1192 led to a 10-fold increase in *in vitro* replication, as previously reported[54] (**Supplementary Figure S13**). Introduction of the change from a serine to an asparagine at position 2414 resulted in approximately a 10-fold increase in replication (**Supplementary Figure S13**). However, introducing a range of other mutations nominally associated with higher viral load produced more variable outcomes. The mutant I1851V replicons show a 10-fold reduced replication compared to the wild type S52 replicon while others showed 2-fold to 3-fold increase. (**Supplementary Figure S13**). No mutations associated with *IFNL4* genotypes other than N2414S led to altered replication phenotypes *in vitro.*

We assessed the relationship between *IFNL4* genotypes, viral amino acids and viral load in genotype 3a isolates by estimating the effect of the CC genotype on the chances of observing a change from the consensus (to non-consensus) using a 2x2 Fisher's exact test (only using data in which the consensus is inferred to be ancestral). These odds ratios were compared against the estimated effect of non-consensus amino acids on viral load (whilst accounting for *IFNL4* genotypes) at sites associated with the *IFNL4* genotypes at a 20% FDR (**Supplementary Figure S14**). We observe a positive relationship (R=0.42, P=0.005) whereby non-consensus amino acids which are increased in frequency in individuals with a CC genotype tend to also associate with increased viral load. The same positive relationship was observed when estimating the effect on viral load in individuals with a CC genotype only (R=0.35, P=0.02) or a non-CC genotype only (R=0.4 and P=0.007). A nominally significant trend was observed when the analysis was done for all variant positions in the viral genome (R=0.099 and P=0.04) (**Supplementary Figure S15)**.

Discussion

Here we report the first systematic analysis of associations between variation in human and HCV genomes in a large patient cohort. Advances in DNA and RNA

sequencing technology and new bioinformatics tools have allowed full-length viral consensus sequences to be obtained in large number of patients for reasonable cost (approximately £100 per sample), as well as host genetic data at millions of directly assayed and imputed polymorphisms (approximately £75 per sample). We apply a fast and simple approach to test for association between host and pathogen variants, using a logistic regression analysis corrected for both human and viral population structures by using the principal components of the genome-wide data as covariates (60 hours for approximately 2500 association studies of 330,000 SNPs). We also applied a contingency table analysis based on the inferred viral amino acid changes since infection. We anticipate that with the reduction in the cost of sequencing and genotyping, and the increasing interest in studying large patient cohorts, analyses of this kind will become a powerful approach to understanding infectious diseases.

We found strong evidence for the adaptive immune system exerting selective pressures on the HCV genome, presumably by preferentially selecting viral mutations that avoid antigenic presentation by the host's HLA proteins. Some of the observed associations are located in experimentally validated viral epitopes[20], however others have not been described experimentally and most likely represent sites of novel T cell escape mutations. Assuming that our analysis removes biases associated with population structure and incorrect ancestral inference, 5% of the viral amino acids (153/3021) are associated with HLA alleles (at 20% FDR). These data highlight the importance of the adaptive immune system in driving viral evolution, and serves as a map of the targets of T-cell based immunity along the HCV genome which can aid vaccine design and development[20].

In addition to the HLA, we now show that *IFNL4* activity may significantly shape the HCV viral genome. Previous studies have shown the "favourable" CC *IFNL4* genotype increases the chances of spontaneous resolution and interferon-based treatment success[2–6]. Prokunina-Olsson et al. have shown that the "favourable" CC *IFNL4* genotype abolishes the expression of *IFNL4*[46], whereas in individuals with the "unfavourable" non-CC *IFNL4* genotype, *IFNL4* expression is maintained leading to the downstream up-regulation of hepatic ISGs expression via the JAK-STAT pathway[49]. The expression of ISGs has been shown to render the host less susceptible to exogenous INF α/γ stimulation and is associated with more infected cells in the liver[56]. To date it has been presumed that this fully explains why patients with specific *IFNL4* genotypes have differential outcomes during primary infection or

with drug therapy[56,57]. Our analysis adds to this story, with the observation that *IFNL4* variants also impact significantly on the HCV viral genome at multiple amino acid sites. Indeed, these were the strongest footprinting signals in our systematic analysis outside HLA region. The most significant association was at position 2570 in the viral NS5B protein; an association that we replicated in an independent cohort infected with HCV genotype 2 or 3. An association between *IFNL4* and amino-acid variability in NS5B protein, has previously been reported in a candidate gene study[38] in a single source infection cohort. *IFNL4* has also been associated with a viral SNP associated with DAA resistance in HCV genotype 1 infection[58] although this association has not been replicated in our HCV genotype 3 cohort[59]. However, the broader impact of interferon lambda host genes that are associated with, and potentially select for, specific viral variants has not been previously recognised.

As far as we are aware, we also report the first association between a single virus amino acid variant (serine vs. non-serine at position 2414 in NS5A) and HCV viral load. HCV viral load is an important and clinically relevant parameter since patients with higher HCV viral loads have lower response rates to IFN and DAA based therapy[60] (independent of *IFNL4* status). Paradoxically, the "favourable" *IFNL4* variants have also been associated with both an increase in disease progression[61] and high viral load[3,62]; our data shows that site 2414 is one of the 11 sites that are putatively associated with *IFNL4* genotypes. Further, our data shows that a decrease in viral load was *only* observed in those patients with the non-CC genotype whose virus carried the serine amino acid at site 2414 in NS5A protein. Since the 2414 S variant is found in 85% of non-CC patients (compared to 67% of CC patients), this interplay between host and viral genes helps explain the previous observation that non-CC patients have a lower HCV viral load **(Figure 6)**. Our in vitro data from a genotype 3 replicon assay shows that a change from a serine to asparagine is associated with an increase in RNA replication and perhaps hyper-phosphorylation[63], which is a negative regulator of virus replication.

Our interpretation of the data (**Figure 6**) is that expression of IFNL4 by the "unfavourable" non-CC genotype leads to the activation of additional components of the immune response, likely driven by ISGs, which interact directly with specific amino acids in the viral genome (most notably amino acid 2414 in NS5A which has a significant impact on viral load (**Figures 2b, 5 and S14**)). Our data suggest that this also leads to an overall increase in the strength of purifying selection (decrease in dN/dS, **Figure 4**), and together this leads to lower viral load. However viruses that

establish chronic infections in non-CC patients have evolved to survive in the more hostile environment (for example mutating the serine at amino acid 2414 of NS5A), which makes them less likely to respond to interferon-based therapy. In contrast, our analysis suggests that the “favourable” CC genotype and the inactivation of *IFNL4* gene disables components of the immune response (therefore removing the effect of amino acid 2414 in NS5A on viral load), which leads to a reduced level of purifying selection (**Figure 4**). It is possible that this then permits a range of mutations which confer higher replicative fitness and therefore higher viral load **(Figure S14),** but these viruses are more susceptible to interferon-based treatments. At the population level, we would expect a balance in the relative contribution of these mechanisms as viruses move between individuals with CC and non-CC genotypes. Our results make the prediction that the outcome of a new infection will be dependent on both the HLA alleles and the *IFNL4* genotype of the patient who is the source of the new infection. Further analysis is required to fully understand the impact of CC and non-CC genotypes on the different components of the immune system and to establish their clinical relevance before, during and after infection.

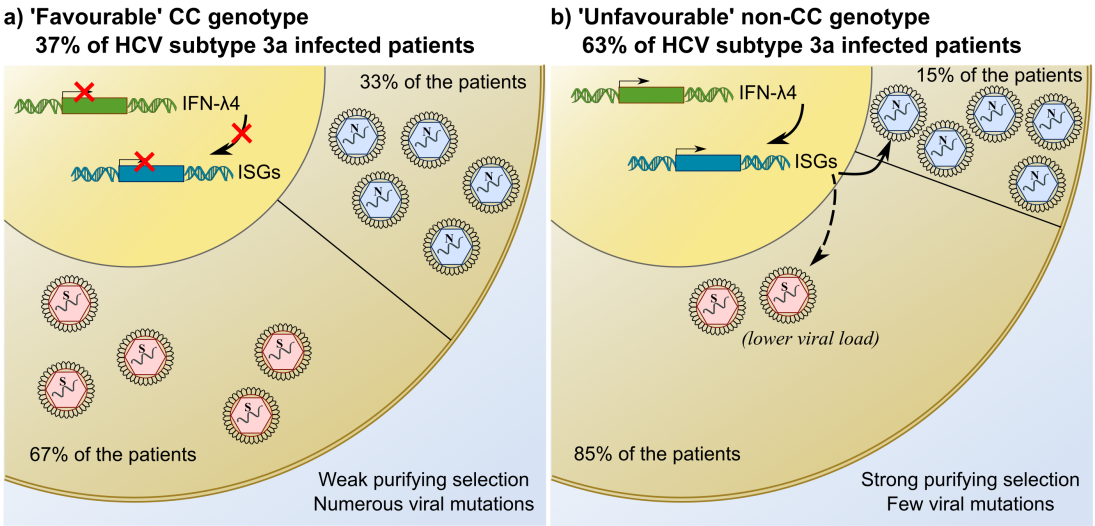


Figure 6: Overview of the observations relating to the interplay between innate immune response and the viral genome in hepatitis C virus (HCV) control. a) Infected individuals with *IFNL4* CC genotype show high rates of spontaneous and treatment-induced clearance of HCV. *IFNL4* is not expressed, which in turn induces a weaker and possibly differential interferon-stimulated genes (ISG) expression. The host environment is associated with weaker purifying selection and allows viral mutations associated with a better replicative fitness to accumulate, leading to higher viral load. b) Infected individuals with *IFNL4* non-CC genotype, have lower rates of spontaneous and treatment-induced clearance of HCV. *IFNL4* is expressed and induces ISGs that collectively establish an antiviral state hostile to viral replication. This hostile environment induces a high selective pressure and fewer viral mutations can accumulate.

In conclusion, we provide a comprehensive genome-to-genome analysis in chronic HCV infections. Using this genome-wide, hypothesis-free, approach we show that the host's HLA alleles leave multiple footprints in the HCV genome and that the host's innate immune environment also influences the presence of amino acid polymorphism in the virus, both at specific loci, and genome-wide. We observe a common viral amino-acid residue that is associated with HCV viral load only in patients with the "unfavourable" non-CC *IFNL4* genotype. These observations suggest that the innate and adaptive immune system jointly impact on HCV genome evolution and likely together determine the establishment of infection and its control over time. The new insights into the biological mechanisms that drive HCV evolution *in vivo*, and the identification of specific interactions between viral and host polymorphisms, are relevant for future approaches to treatment stratification and vaccine development.

Materials & Correspondence.

Correspondence and material requests should be addressed to Chris Spencer (chris.spencer@well.ox.ac.uk) or by contacting STOP-HCV http://www.stop-hcv.ox.ac.uk/contact .

Author contributions

M.A.A and V.P contributed equally; E.B and C.C.A.S jointly supervised research; M.A.A, V.P, E.B and C.C.A.S conceived and designed the experiments; M.A.A, V.P, C.I, A.M, A.V, N.C, I.B, A.T and P.P performed the experiments; M.A.A, V.P, A.M, N.C, I.B and C.C.A.S performed statistical analysis; M.A.A, V.P, A.M, P.K, E.B, C.C.A.S analysed the data; C.I, G.N, A.V, D.B, G.M, A.T, P.P, J.F and J.M contributed reagents/materials/analysis tools; M.A.A, V.P, J.F, G.C, G.R.F, E.H, J.M, P.M, R.B, P.K, E.B and C.C.A.S wrote the paper.

Acknowledgements

The authors would like to thank Gilead Sciences for the provision of samples and data from the BOSON clinical study for use in these analyses. The authors would also like to thank HCV Research UK (funded by the Medical Research Foundation) for their assistance in handling and coordinating the release of samples for these analyses.

This work was funded by a grant from the Medical Research Council (MR/K01532X/1 – STOP-HCV Consortium). The work was supported by Core funding to the Wellcome Trust Centre for Human Genetics provided by the Wellcome Trust (090532/Z/09/Z). E.B is funded by the MRC as an MRC Senior clinical fellow with additional support from the Oxford NHIR BRC and the Oxford Martin School. A.M is funded by the Oxford Martin School. G.C is funded by the BRC of Imperial College NHS Trust. P.K is funded by the Oxford Martin School, NIHR Biomedical Research Centre, Oxford, by the Wellcome Trust (091663MA) and NIH (U19AI082630). C.C.A.S is funded by the Wellcome Trust (097364/Z/11/Z). G.M is founded by the Wellcome Trust grant 100956/Z/13/Z.

Conflicts of interest

The authors disclose the following: G.R.F: Grants Consulting and Speaker/Advisory Board: AbbVie, Alcura, Bristol-Myers Squibb, Gilead, Janssen, GlaxoSmithKline,

Merck, Roche, Springbank, Idenix, Tekmira, Novartis. G.M. is a partner in Peptide Groove LLP, which commercializes HLA*IMP.

Online Methods

Patients and sample

Plasma and DNA samples came from patients enrolled in the Boson study. Boson study is a phase 3 randomized open-label trial to determine the efficacy and safety of sofosbuvir with and without pegylated-interferon-alfa, in treatment-experienced patients with cirrhosis and hepatitis C virus (HCV) genotype 2 infection and treatment-naïve or -experienced patients with HCV genotype 3 infection[30]. All patients provided written informed consent before undertaking any study-related procedures. The study protocol was approved by each institution's review board or ethics committee before study initiation. The study was conducted in accordance with the International Conference on Harmonization Good Clinical Practice Guidelines and the Declaration of Helsinki. The study reported here is not a clinical trial, but is based on the analysis of patients from a clinical trial registration number: NCT01962441. Sample sizes were determined by the available data. All samples for which both viral sequencing and host genetics were available were included in the final analysis unless otherwise specified.

Host genotyping and imputation

Informed consent for host genetic analysis was obtained from 567 patients. Genotyping was performed using Affymetrix UK Biobank arrays. We imputed the MHC class I loci *HLA-A*, *HLA-B*, *HLA-C* and class II loci *HLA-DQA1*, *HLA-DQB1*, *HLA-DPB1*, *HLA-DRB1*, *HLA-DRB3*, *HLA-DRB4*, *HLA-DRB5* using HLA*IMP:02[27] accessed 22 March 2015. HLA amino acids were also imputed by SNP2HLA[64] using the T1DGC as the reference panel, which contains 5225 unrelated individuals (10,450 haplotypes). Logistic regression using posterior genotype probabilities (allele dosages) for each HLA allele from SNP2HLA were carried out using PLINK2[32] (https://www.cog-genomics.org/plink2).

Virus sequencing

Sample collection and preparation

RNA was isolated from 500µl plasma using the NucliSENS magnetic extraction system (bioMerieux) and collected in 30µl of kit elution buffer for storage in aliquots at -80°C.

Sequencing library construction, enrichment and sequencing

Libraries were prepared for Illumina sequencing using the NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (New England Biolabs) with 5µl sample (maximum 10ng total RNA) and previously published modifications of the manufacturer's guidelines (v2.0)[23], briefly: fragmentation for 5 minutes at 94°C, omission of Actinomycin D at first-strand reverse transcription, library amplification for 18 PCR cycles using custom indexed primers[65] and post-PCR clean-up with 0.85× volume Ampure XP (Beckman Coulter).

Libraries were quantified using Quant-iT™ PicoGreen® dsDNA Assay Kit (Invitrogen) and analysed using Agilent TapeStation with D1K High Sensitivity kit (Agilent) for equimolar pooling, then re-normalized by qPCR using the KAPA SYBR® FAST qPCR Kit (Kapa Biosystems) for sequencing. A 500ng aliquot of the pooled library was enriched using the xGen® Lockdown® protocol from IDT (Rapid Protocol for DNA Probe Hybridization and Target Capture Using an Illumina TruSeq® or Ion Torrent® Library (v1.0), Integrated DNA Technologies) with equimolar-pooled 120nt DNA oligonucleotide probes (IDT) followed by a 12-cycle, modified, on-bead, post-enrichment PCR re-amplification. The cleaned post-enrichment ve-Seq library was normalized with the aid of qPCR and sequenced with 151b paired-end reads on a single run of the Illumina MiSeq using v2 chemistry.

Sequence data analysis

De-multiplexed sequence read-pairs were trimmed of low-quality bases using QUASR v7.0120[66] and adapter sequences with CutAdapt version 1.7.1[67] and subsequently discarded if either read had less than 50b remaining sequence or if both reads matched the human reference sequence using Bowtie version 2.2.4[68]. The remaining read pool was screened against a BLASTn database containing 165 HCV genomes[69] covering its diversity both to choose an appropriate reference and to select those reads which formed a majority population for de novo assembly with Vicuna v1.3[70] and finishing with V-FAT v1.0 (http://www.broadinstitute.org/scientific-community/science/projects/viral-genomics/v-fat). Population consensus sequence at each site is defined as the most common variant at that site among all the patients.

Phylogenetic and ancestral sequence reconstruction

Whole genome viral consensus sequences for each patient were aligned using MAFFT[71] with default settings. This alignment was used to create a maximum likelihood tree using RAxML[72], assuming a general time reversible model of nucleotide substitution under the gamma model of rate heterogeneity. The resulting

tree was rooted at midpoint. Maximum likelihood ancestral sequence reconstruction was performed using RAxML[72] with the maximum likelihood tree and HCV polyprotein sequences as input.

Association analysis

To test for association between human SNPs and HCV amino acids at genome-to-genome level, we performed logistic regression using PLINK2[32] (https://www.cog-genomics.org/plink2) adjusted for the human population structure (three first PCs assessed using EIGENSOFT v3.0[73]) and the virus population structure (10 first PCs). For the viral data PCA was performed on the nucleotide data. Tri and quad-allelic sites were converted to binary variables and the amino acid frequencies were standardised to have mean zero and unit variance. MATLAB (Release 2015a, The MathWorks) was used to perform the PCA using the singular value decomposition function.

To test for association between imputed HLA alleles and HCV amino acids, we used Fisher's exact test, correcting for the virus population structure as described in Bhatacharaya *et al.*[42]. We used the inferred ancestral amino acid to construct a 2x4 contingency table where rows denote presence or absence of a host HLA allele, and the columns denote changes to and away from a specific amino acid in viruses with and without the amino acid inferred to be ancestral. To test for association between *IFNL4* SNP rs12979860 genotypes and HCV amino acids, we used the same Fisher's exact test with a dominant model for *IFNL4* rs12979860 by encoding genotypes as CC and non-CC.

Permutation was used to estimate the FDR for association tests that used Fisher's exact test. The labels (presence and absence of each HLA allele or the *INFL4* genotypes) were randomly permuted 500 times and in each case the p-values of the associations were calculated. For each threshold t, the expectation of the number of false positives was estimated from result of the permutation tests. FDR for threshold t was then defined as the expected number of false positives divided by the observed number of significant associations in the actual data at threshold t.

To test for enrichment of association signals in epitope regions, viral proteins or with a specific HLA allele we used Fisher's exact test. Each site is either within a target region (epitope regions or a specific viral protein) or not and the most associated test with it, is significant or not with an FDR of 5%. The resulting contingency table was

tested using Fisher's exact test to assess enrichment or depletion of signals of association.

To assess the relationship between rates of escape and reversion in HLA presentation, we estimated the odds ratio for each 2x2 sub-table used in the Fisher's exact test. This was done only for viral sites associated with HLA alleles at 20% FDR and which there were sufficient observations in both tables to estimate the odds ratio where confidence interval did not go to infinity. Pearson's correlation coefficient was used to assess the relationship between the $\log_{10}$(OR) of escape and reversion.

To test for enrichment of viral amino acid associations with host *IFNL4* genotypes in viral proteins, the null distribution of number of association in each protein was estimated using 10,000 permutations of *IFNL4* labels and performing the same tests. The estimated null distribution of number of associations for each viral protein was compared to the observed number of associations in the data to test for enrichment or depletion of number of associations. To test if HVR1, HVR2, HVR3, ISDR and PKR-BD regions show differences in the number of changes away from the population consensus in CC and non-CC hosts, we used a Poisson regression. In each individual and each locus we determined the number of differences to the population consensus. We then estimated the effect of *INFL4* genotypes on the mean number of differences to the population consensus using Poisson regression. The same procedure was used to test if the total number of differences across the whole poly-protein relative to the population consensus was influenced by *INFL4* genotypes.

To estimate the rate of synonymous and non-synonymous mutations, we used dndsml function from MATLAB (Release 2015a, The MathWorks) that uses Goldman and Yang's method. It estimates (using maximum likelihood) an explicit model for codon substitution that takes into account transition/transversion rate bias and base/codon frequency bias. Then it uses the model to correct synonymous and non-synonymous counts to account for multiple substitutions at the same site. To estimate dN and dS, each sequence was compared to the population consensus which indicates the most common nucleotide observed in our data set at each position along the genome.

To determine whether *IFNL4* genotypes impact on HLA alleles presentation of epitopes, logistic regression with interaction term was used. The outcome for

individuals was whether on the terminal branches of the tree a specific had changed amino or not. We tested for interaction between presence and absence of the associated HLA allele and *IFNL4* genotypes for all combinations of HLA alleles and viral sites associated at a 20% FDR. In addition, we tested for an overall effect of *INFL4* genotypes on HLA alleles' presentation of epitopes. We used our 2x4 contingency tables and the odds ratios estimated from the 2x2 sub-tables to infer the antigenic amino acids. If the odds ratio indicates that in individuals with HLA allele present, the "X ancestral amino acid -> any other amino acid" element is enriched relative to individuals without the HLA allele then we assume the X amino acid is the antigenic amino acid and escape occurs away from X to any other amino acid. For these cases, we count how many of "X ancestral amino acid -> any other amino acid" occur in CC and non-CC individuals across all combinations of HLA alleles and viral sites associated at 20% FDR. If *IFNL4* genotypes have no impact on the HLA presentation (null hypothesis), then the mean number of escape mutations in CC and non-CC hosts should be proportional to the frequency of hosts with CC and non-CC genotypes (null distribution is binomial with parameters n equal to the total number of observed escape mutations and p equal to the proportion of CC hosts).

We used linear regression in PLINK2[32] (https://www.cog-genomics.org/plink2) to test for association between human SNPs and $\log_{10}$ transformed pre-treatment viral load (PTVL) including sex and first three PCs of the host genome as covariates. We used linear regression to test for association between HCV amino acids (with a minimal count of 10 at each site) and viral load. We used linear regression in R (version 3.2.4 (2016-03-10))[74] to analyse the interaction between *IFNL4* genotypes and amino acids at viral site 2414 and to quantify their impact on viral load.

For all viral sites associated with *IFNL4* genotypes at 20% FDR, we assessed if there is a relationship between effect size of non-consensus amino acids on viral load and effect size of *IFNL4* genotypes and changes to non-consensus amino acids. We estimated the odds ratio of enrichment of changes away from the consensus amino acid on the terminal branches of the virus tree. We also estimated the effect size of non-consensus amino acids on the $\log_{10}$ of viral load using a linear regression with *IFNL4* genotype as a covariate. Additionally, we estimated the effect size of non-consensus amino acid on the $\log_{10}$ viral load in CC and non-CC hosts. We used Pearson's correlation coefficient to measure the strength of relationship between the effect size of non-consensus amino acids on viral load and log of odds ratio of enrichment of non-consensus amino acid changes in CC hosts.

Data and Code Availability

The consensus viral sequences will be made available via GENBANK (http://www.ncbi.nlm.nih.gov/genbank/), and the human genotype data will be made available via EGA (https://www.ebi.ac.uk/ega/) through the STOP-HCV data access committee http://www.stop-hcv.ox.ac.uk. R and MATLAB code used to generate the results and figures from the primary analyses described above are available from the authors on request.

Replication study

To replicate the *IFNL4* SNP rs12979860 results, we ran the association analysis on an independent HCV infected population that was recruited to the FISSION, FUSION and POSITRON phase 3 clinical studies[75,76] . Paired human genome-wide genotyping and HCV sanger sequencing data for NS5B amplicon were obtained from DNA and plasma samples collected from 360 Caucasian patients chronically infected with HCV genotype 2 (N=153) or 3 (N=208). We searched for association between the *IFNL4* SNP rs12979860 and viral position 2570 using logistic regression and a binary vector as the outcome which indicated the presence or absence of amino acid V for all samples at position 2570. To prevent spurious associations due to host and viral stratification, we included human principal components and viral genotype as covariates.

Replicon assay

Cell Culture

Huh7.5-Sec14L2 cells, previously reported[53] were grown in Dulbecco's modified Eagle's medium (DMEM, Life Technologies) supplemented with 10% fetal calf serum, 2 mM L-glutamine, 100 U/ml penicillin, 100 0 enicillin, 100 100 lmM HEPES and 0.1M nonessential amino acids as described[77]. Huh7.5 cells are heterozygous CT for the *IFNL4* rs12979860 SNP[78].

HCV Mutant Replicons

The enhanced version of the subgenomic replicon of genotype 3a strain S52, lacking of neomycin resistance gene, has been previously described[53]. Site-specific mutations were introduced using QuikChange II XL Site-Directed Mutagenesis Kit (Agilent Technologies) following manufacturer instructions and confirmed by direct sequencing. HCV mutant plasmids were linearized by *XbaI* digestion (New England Biolabs; NEB), mung-bean treated (NEB) and purified. Linearized DNA was then used as template for *in vitro* RNA transcription (IVT) (Megascript T7, Life

Technologies) according to manufacturer protocol. Finally, IVT RNA has been DNAse treated, purified and stored at -80°C.

Electroporation and Luciferase Detection

For electroporation, cells were counted and then washed twice in ice-cold PBS. Typically, for each mutant $4x10^6$ cells were mixed with 1 µg of replicon RNA in a 4mm cuvette and electroporated in the Gene Pulser Xcell (Bio-Rad) at 250 V, 950 µF using exponential decay setting. Cells were immediately recovered in pre-warmed complete DMEM, seeded in a 24-well plate and incubated at 37°C. After 5, 24, 48 or 72 hours, medium was removed and cells lysed with Glo Lysis Buffer (Promega). Cell lysates were then transferred in a white 96-well plate (Corning) and the luciferase expression was quantitated in a luminometer (GloMax 96 Microplate Luminometer, Promega) using Bright-Glo assay system (Promega).

References

1. Mohd Hanafiah, K., Groeger, J., Flaxman, A. D. & Wiersma, S. T. Global epidemiology of hepatitis C virus infection: new estimates of age-specific antibody to HCV seroprevalence. *Hepatology* **57,** 1333–1342 (2013).
2. Thomas, D. L. *et al.* Genetic variation in IL28B and spontaneous clearance of hepatitis C virus. *Nature* **461,** 798–801 (2009).
3. Ge, D. *et al.* Genetic variation in IL28B predicts hepatitis C treatment-induced viral clearance. *Nature* **461,** 399–401 (2009).
4. Rauch, A. *et al.* Genetic variation in IL28B is associated with chronic hepatitis C and treatment failure: a genome-wide association study. *Gastroenterology* **138,** 1338–45, 1345–7 (2010).
5. Suppiah, V. *et al.* IL28B is associated with response to chronic hepatitis C interferon-alpha and ribavirin therapy. *Nat. Genet.* **41,** 1100–1104 (2009).
6. Tanaka, Y. *et al.* Genome-wide association of IL28B with response to pegylated interferon-alpha and ribavirin therapy for chronic hepatitis C. *Nat. Genet.* **41,** 1105–1109 (2009).
7. Enomoto, N. *et al.* Mutations in the Nonstructural Protein 5a Gene and Response to Interferon in Patients with Chronic Hepatitis C Virus 1b Infection. *N. Engl. J. Med.* **334,** 77–82 (1996).
8. Pascu, M. *et al.* Sustained virological response in hepatitis C virus type 1b infected patients is predicted by the number of mutations within the NS5A-ISDR: a meta-analysis focused on geographical differences. *Gut* **53,** 1345–1351 (2004).
9. Halfon, P. & Locarnini, S. Hepatitis C virus resistance to protease inhibitors. *J. Hepatol.* **55,** 192–206 (2011).
10. Halfon, P. & Sarrazin, C. Future treatment of chronic hepatitis C with direct acting antivirals: Is resistance important? *Liver International* **32,** 79–87 (2012).
11. Ahmed, A. & Felmlee, D. J. Mechanisms of hepatitis C viral resistance to direct acting antivirals. *Viruses* **7,** 6716–6729 (2015).
12. Sarrazin, C. *et al.* Prevalence of Resistance-Associated Substitutions in HCV NS5A, NS5B, or NS3 and Outcomes of Treatment with Ledipasvir and Sofosbuvir. *Gastroenterology* (2016). doi:10.1053/j.gastro.2016.06.002
13. Messina, J. P. *et al.* Global distribution and prevalence of hepatitis C virus

genotypes. *Hepatology* 77–87 (2014). doi:10.1002/hep.27259
14. Pol, S., Vallet-Pichard, A. & Corouge, M. Treatment of hepatitis C virus genotype 3-infection. *Liver Int.* **34 Suppl 1,** 18–23 (2014).
15. Ampuero, J., Romero-Gómez, M. & Reddy, K. R. Review article: HCV genotype 3 – the new treatment challenge. *Aliment. Pharmacol. Ther.* **39,** 686–698 (2014).
16. Fitzmaurice, K. *et al.* Molecular footprints reveal the impact of the protective HLA-A*03 allele in hepatitis C virus infection. *Gut* **60,** 1563–1571 (2011).
17. Neumann-Haefelin, C. *et al.* Human leukocyte antigen B27 selects for rare escape mutations that significantly impair hepatitis C virus replication and require compensatory mutations. *Hepatology* **54,** 1157–1166 (2011).
18. Heim, M. H. & Thimme, R. Innate and adaptive immune responses in HCV infections. *J. Hepatol.* **61,** S14–S25 (2014).
19. Swadling, L. *et al.* A human vaccine strategy based on chimpanzee adenoviral and MVA vectors that primes, boosts, and sustains functional HCV-specific T cell memory. *Sci. Transl. Med.* **6,** 261ra153 (2014).
20. von Delft, A. *et al.* The broad assessment of HCV genotypes 1 and 3 antigenic targets reveals limited cross-reactivity with implications for vaccine design. *Gut* **65,** 112–123 (2016).
21. Simmonds, P. Genetic diversity and evolution of hepatitis C virus--15 years on. *J. Gen. Virol.* **85,** 3173–3188 (2004).
22. Lauck, M. *et al.* Analysis of hepatitis C virus intrahost diversity across the coding region by ultradeep pyrosequencing. *J. Virol.* **86,** 3952–3960 (2012).
23. Bonsall, D. *et al.* ve-SEQ: Robust, unbiased enrichment for streamlined detection and whole-genome sequencing of HCV and other highly diverse pathogens. *F1000Research* **4,** 1062 (2015).
24. Thomson, E. *et al.* Comparison of next generation sequencing technologies for the comprehensive assessment of full-length hepatitis C viral genomes. *J. Clin. Microbiol.* (2016). doi:10.1128/JCM.00330-16
25. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
26. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genomics Hum. Genet.* **10,** 387–406 (2009).
27. Dilthey, A. *et al.* Multi-Population Classical HLA Type Imputation. *PLoS Comput. Biol.* **9,** (2013).
28. Zheng, X. *et al.* HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J.* **14,** 192–200 (2014).
29. Bartha, I. *et al.* A genome-to-genome analysis of associations between human genetic variation, HIV-1 sequence diversity, and viral control. *Elife* **2,** e01123 (2013).
30. Foster, G. R. *et al.* Efficacy of sofosbuvir plus ribavirin with or without peginterferon-alfa in patients with hepatitis C virus genotype 3 infection and treatment-experienced patients with cirrhosis and hepatitis C virus genotype 2 infection. *Gastroenterology* **149,** 1462–1470 (2015).
31. Gonzalez-Galarza, F. F. *et al.* Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.* **43,** D784–D788 (2015).
32. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4,** 7 (2015).
33. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet. Epidemiol.* **32,** 381–5 (2008).
34. Johnson, R. C. *et al.* Accounting for multiple comparisons in a genome-wide association study (GWAS). *BMC Genomics* **11,** 724 (2010).
35. de Bakker, P. I. W. *et al.* A high-resolution HLA and SNP haplotype map for

disease association studies in the extended human MHC. *Nat. Genet.* **38,** 1166–1172 (2006).
36. Malkki, M., Single, R., Carrington, M., Thomson, G. & Petersdorf, E. MHC microsatellite diversity and linkage disequilibrium among common HLA-A, HLA-B, DRB1 haplotypes: implications for unrelated donor hematopoietic transplantation and disease association studies. *Tissue Antigens* **66,** 114–124 (2005).
37. Ruhl, M. *et al.* CD8+ T-cell response promotes evolution of hepatitis c virus nonstructural proteins. *Gastroenterology* **140,** 2064–2073 (2011).
38. Merani, S. *et al.* Effect of immune pressure on hepatitis C virus evolution: insights from a single-source outbreak. *Hepatology* **53,** 396–405 (2011).
39. Rauch, A. *et al.* Divergent adaptation of hepatitis C virus genotypes 1 and 3 to human leukocyte antigen-restricted immune pressure. *Hepatology* **50,** 1017–1029 (2009).
40. Gaudieri, S. *et al.* Evidence of Viral Adaptation to HLA Class I-Restricted Immune Pressure in Chronic Hepatitis C Virus Infection. *J. Virol.* **80,** 11094–11104 (2006).
41. Pagel, M. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. *Syst. Biol.* **48,** 612–622 (1999).
42. Bhattacharya, T. *et al.* Founder effects in the assessment of HIV polymorphisms and HLA allele associations. *Science* **315,** 1583–1586 (2007).
43. Westfall, P. H. & Young, S. S. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. (Wiley, 1993).
44. Meinshausen, N., Maathuis, M. H. & Bühlmann, P. Asymptotic optimality of the Westfall–Young permutation procedure for multiple testing under dependence. *Ann. Stat.* **39,** 3369–3391 (2011).
45. Neumann-Haefelin, C. *et al.* Analysis of the evolutionary forces in an immunodominant CD8 epitope in hepatitis C virus at a population level. *J. Virol.* **82,** 3438–3451 (2008).
46. Bronke, C. *et al.* HIV escape mutations occur preferentially at HLA-binding sites of CD8 T-cell epitopes. *AIDS* **27,** 899–905 (2013).
47. Patin, E. *et al.* Genome-wide association study identifies variants associated with progression of liver fibrosis from HCV infection. *Gastroenterology* **143,** 1212–1244 (2012).
48. Prokunina-Olsson, L. *et al.* A variant upstream of IFNL3 (IL28B) creating a new interferon gene IFNL4 is associated with impaired clearance of hepatitis C virus. *Nat. Genet.* **45,** 164–171 (2013).
49. Terczyńska-Dyla, E. *et al.* Reduced IFNλ4 activity is associated with improved HCV clearance and reduced expression of interferon-stimulated genes. *Nat. Commun.* **5,** 5699 (2014).
50. Wong, M.-T. & Chen, S. S.-L. Emerging roles of interferon-stimulated genes in the innate immune response to hepatitis C virus infection. *Cell. Mol. Immunol.* (2014). doi:10.1038/cmi.2014.127
51. Saito, T. *et al.* Sequence analysis of PePHD within HCV E2 region and correlation with resistance of interferon therapy in Japanese patients infected with HCV genotypes 2a and 2b. *Am. J. Gastroenterol.* **98,** 1377–1383 (2003).
52. Macquillan, G. C. *et al.* Does sequencing the PKRBD of hepatitis C virus NS5A predict therapeutic response to combination therapy in an Australian population? *J. Gastroenterol. Hepatol.* **19,** 551–557 (2004).
53. Witteveldt, J., Martin-Gans, M. & Simmonds, P. Enhancement of the Replication of Hepatitis C Virus Replicons of Genotypes 1 to 4 by Manipulation of CpG and UpA Dinucleotide Frequencies and Use of Cell Lines Expressing SECL14L2 for Antiviral Resistance Testing. *Antimicrob. Agents Chemother.* **60,** 2981–2992 (2016).

54. Yu, M. *et al.* In vitro efficacy of approved and experimental antivirals against novel genotype 3 hepatitis C virus subgenomic replicons. *Antiviral Res.* **100,** 439–445 (2013).
55. Bibert, S. *et al.* IL28B expression depends on a novel TT/-G polymorphism which improves HCV clearance prediction. *J. Exp. Med.* **210,** 1109–1116 (2013).
56. Sheahan, T. *et al.* Interferon Lambda Alleles Predict Innate Antiviral Immune Responses and Hepatitis C Virus Permissiveness. *Cell Host Microbe* **15,** 190–202 (2014).
57. Ferraris, P. *et al.* Cellular Mechanism for Impaired Hepatitis C Virus Clearance by Interferon Associated with IFNL3 Gene Polymorphisms Relates to Intrahepatic Interferon-λ Expression. *Am. J. Pathol.* **186,** 938–951 (2016).
58. Peiffer, K.-H. *et al.* Interferon lambda 4 genotypes and resistance-associated variants in patients infected with hepatitis C virus genotypes 1 and 3. *Hepatology* **63,** 63–73 (2016).
59. Pedergnana, V. *et al.* Interferon Lambda 4 variant rs12979860 is not associated with RAV NS5A Y93H in Hepatitis C Virus Genotype 3a. *Hepatology* (2016). doi:10.1002/hep.28533
60. McHutchison, J. G. *et al.* Peginterferon alfa-2b or alfa-2a with ribavirin for treatment of hepatitis C infection. *N. Engl. J. Med.* **361,** 580–593 (2009).
61. Bochud, P.-Y. *et al.* IL28B alleles associated with poor hepatitis C virus (HCV) clearance protect against inflammation and fibrosis in patients infected with non-1 HCV genotypes. *Hepatology* **55,** 384–394 (2012).
62. Thompson, A. J. *et al.* Interleukin-28B Polymorphism Improves Viral Kinetics and Is the Strongest Pretreatment Predictor of Sustained Virologic Response in Genotype 1 Hepatitis C Virus. *Gastroenterology* **139,** (2010).
63. Tellinghuisen, T. L., Foss, K. L. & Treadaway, J. Regulation of hepatitis C virion production via phosphorylation of the NS5A protein. *PLoS Pathog.* **4,** e1000032 (2008).
64. Jia, X. *et al.* Imputing Amino Acid Polymorphisms in Human Leukocyte Antigens. *PLoS One* **8,** (2013).
65. Lamble, S. *et al.* Improved workflows for high throughput library preparation using the transposome-based Nextera system. *BMC Biotechnol.* **13,** 104 (2013).
66. Gaidatzis, D., Lerch, A., Hahne, F. & Stadler, M. B. QuasR: Quantification and annotation of short reads in R. *Bioinformatics* **31,** 1130–1132 (2015).
67. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10 (2011).
68. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9,** 357–359 (2012).
69. Smith, D. B. *et al.* Expanded classification of hepatitis C virus into 7 genotypes and 67 subtypes: updated criteria and genotype assignment web resource. *Hepatology* **59,** 318–327 (2014).
70. Yang, X. *et al.* De novo assembly of highly diverse viral populations. *BMC Genomics* **13,** 475 (2012).
71. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol. Biol. Evol.* **30,** 772–780 (2013).
72. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
73. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38,** 904–909 (2006).
74. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

75. Jacobson, I. M. *et al.* Sofosbuvir for Hepatitis C Genotype 2 or 3 in Patients without Treatment Options. *N. Engl. J. Med.* **368,** 1867–1877 (2013).
76. Lawitz, E. *et al.* Sofosbuvir for previously untreated chronic hepatitis C infection. *N. Engl. J. Med.* **368,** 1878–87 (2013).
77. Magri, A. *et al.* Rethinking the old antiviral drug moroxydine: Discovery of novel analogues as anti-hepatitis C virus (HCV) agents. *Bioorg. Med. Chem. Lett.* **25,** 5372–5376 (2015).
78. Rojas, Á. *et al.* Hepatitis C virus infection alters lipid metabolism depending on IL28B polymorphism and viral genotype and modulates gene expression in vivo and in vitro. *J. Viral Hepat.* **21,** 19–24 (2014).

Table 1. Associations between HLA alleles and viral amino acids at a 5% false discovery rate. For each combination of HLA allele and viral site, only the most significant associated amino acid is reported. The first amino acid at "variable amino acids at this site" column indicates the most common amino acid at the site in our data.

HLA allele	HCV amino acid position	Viral protein	Variable amino acids at this site	Associated amino acid	P	q value	In a known epitope?
B*07:02	42	C	PL	P	$1.55x10^{-07}$	$5.00x10^{-03}$	Yes
C*07:02	42	C	PL	P	$2.27x10^{-07}$	$6.10x10^{-03}$	Yes
C*15:02	561	E2	VTLI	L	$5.70x10^{-07}$	$1.50x10^{-02}$	No
A*31:01	1270	NS3	RKHS	H	$4.21x10^{-10}$	$< 4.00x10^{-04}$	Yes
A*33:03	1282	NS3	NVTSA	T	$1.73x10^{-07}$	$5.26x10^{-03}$	Yes
B*15:01	1290	NS3	KPARS	R	$1.25x10^{-07}$	$4.47x10^{-03}$	Yes
A*68:02	1341	NS3	VA	V	$2.43x10^{-09}$	$4.44x10^{-04}$	No
A*68:02	1344	NS3	TVA	T	$3.68x10^{-10}$	$< 4.00x10^{-04}$	No
B*51:01	1380	NS3	ILV	L	$3.22x10^{-08}$	$1.67x10^{-03}$	Yes
A*01:01	1444	NS3	FY	F	$9.63x10^{-33}$	$< 4.00x10^{-04}$	Yes
A*68:02	1452	NS3	IV	I	$6.98x10^{-09}$	$4.44x10^{-04}$	No
A*30:02	1585	NS3	YF	Y	$1.19x10^{-06}$	$3.28x10^{-02}$	No
B*13:02	1635	NS3	ITVLAF	T	$7.38x10^{-08}$	$3.13x10^{-03}$	No
B*08:01	1646	NS3	MTAVSI	T	$2.89x10^{-12}$	$< 4.00x10^{-04}$	No
C*07:01	1646	NS3	MTAVSI	T	$3.01x10^{-07}$	$7.39x10^{-03}$	No
DQA1*05:01	1646	NS3	MTAVSI	T	$7.24x10^{-07}$	$1.76x10^{-02}$	No
A*02:01	1873	NS4B	LFKICVPRMTA	F	$4.06x10^{-10}$	$< 4.00x10^{-04}$	No
B*38:01	2034	NS5A	STNLPIAQVDKEMGRFW	P	$7.56x10^{-08}$	$3.13x10^{-03}$	Yes
B*18:01	2144	NS5A	ED	E	$1.16x10^{-06}$	$3.28x10^{-02}$	Yes
A*31:01	2510	NS5B	AKQSEMGLTV	A	$4.96x10^{-08}$	$2.14x10^{-03}$	No
A*32:01	2537	NS5B	NDHSYEAI	N	$1.13x10^{-08}$	$7.27x10^{-04}$	Yes
A*32:01	2540	NS5B	RSKHNQC	R	$7.45x10^{-09}$	$6.00x10^{-04}$	Yes
A*02:11	2600	NS5B	QKRSAL	Q	$1.01x10^{-06}$	$3.12x10^{-02}$	Yes
A*26:01	2605	NS5B	EAGVK	G	$2.61x10^{-07}$	$6.64x10^{-03}$	Yes

Variable	
Age in years (Mean (range))	49.8 (19-73)
log_{10} pre treatment viral load (Mean (range))	6.3 (3.9-7.6)
Self reported ethnicity (N (%))	
White	452 (83.2%)
Asian	74 (13.8 %)
Other	16 (2.9%)
Sex (N (%)))	
Female	177 (32.6%)
Male	365 (67.4%)
IFNL4 SNP rs12979860 (N (%))	
CC	206 (38.1%)
TC	264 (48.6%)
TT	72 (13.3%)
HCV genotype	
3a	487 (89.9%)
3 non a	9 (1.6%)
2	46 (8.5%)

Table S1. Demographic and clinical data of the cohort.

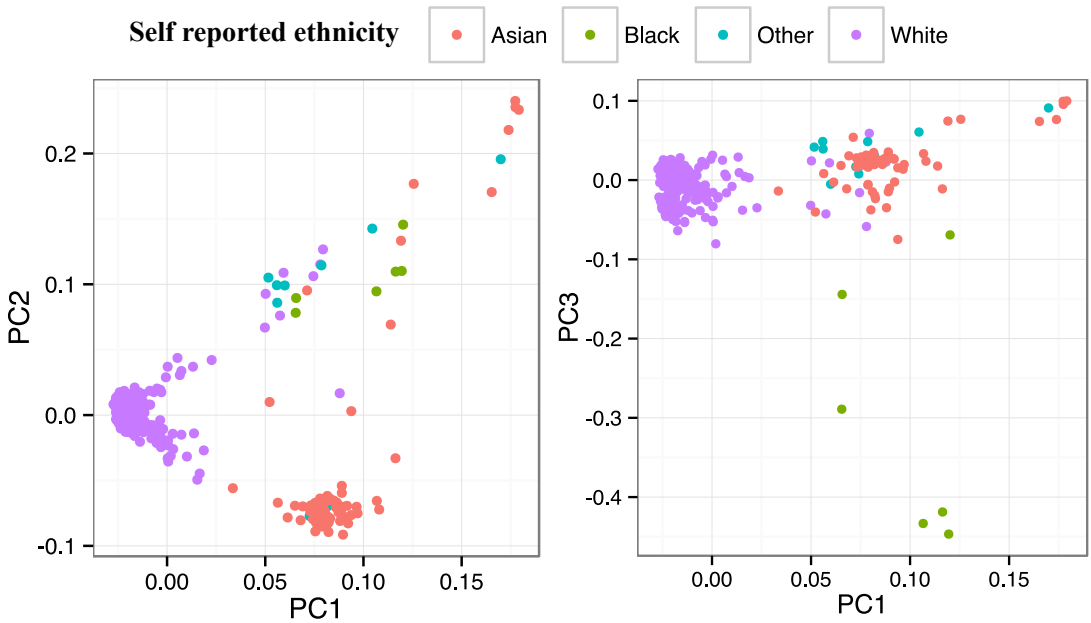


Figure S1. Ancestry stratifies patterns of human genetic variation. PCA plots (1st, 2nd and 3rd components) estimated from human genotypes. Individuals are coloured by self-reported population of origin. Because of the over-representation of individuals from the respective populations, the 1st and 2nd components differentiate South Asian (Pakistan, India) and Western European patients, while the 3rd component differentiates a small number of patients with African ancestry from the other populations.

HLA allele	Asian (%)	White (%)	*P* (Fisher)
DQB1*06:01	19	0.66	1.83E-10
DRB1*15:02	16	0.66	8.49E-09
C*12:02	15	0.88	1.84E-07
B*52:01	14	0.88	1.09E-06
B*40:06	9.5	0.22	6.01E-06
B*13:01	8.1	0	6.48E-06
A*02:01	20	48	7.48E-06
B*07:02	5.4	28	7.64E-06
B*44:02	2.7	22	1.17E-05
A*11:01	34	13	1.96E-05
A*02:11	6.8	0	4.89E-05
C*05:01	2.7	20	5.43E-05
DPA1*02:02	22	6.2	8.48E-05
DRB1*01:01	1.4	16	1.10E-04
DQA1*01:03	30	12	2.37E-04
A*33:03	6.8	0.22	2.61E-04
DRB1*12:02	6.8	0.22	2.61E-04
C*04:03	5.4	0	3.65E-04
C*03:02	6.8	0.44	8.13E-04
DQB1*06:02	11	28	9.12E-04
DRB1*10:01	11	2.2	1.44E-03
C*08:01	5.4	0.22	1.63E-03
C*15:02	14	3.8	1.92E-03
DPA1*01:03	88	97	2.11E-03
A*02:06	4.1	0	2.69E-03
B*15:02	4.1	0	2.69E-03
C*07:01	14	30	3.07E-03
DRB1*14:04	6.8	0.88	3.87E-03
B*51:01	20	8.4	5.43E-03
A*26:01	14	4.6	6.21E-03
C*16:02	5.4	0.66	9.10E-03
DQA1*06:01	5.4	0.66	9.10E-03
C*14:02	9.5	2.7	1.02E-02
C*08:02	0	7.1	1.47E-02
C*02:02	1.4	9.3	1.96E-02
B*27:07	2.7	0	1.96E-02
B*46:01	2.7	0	1.96E-02
DQA1*03:01	18	30	2.63E-02
DRB1*04:01	5.4	15	2.71E-02
A*03:01	15	27	2.99E-02
DRB3*01:01	15	27	2.99E-02
B*27:05	1.4	8.6	3.01E-02
A*31:01	12	5.3	3.56E-02
B*14:02	0	5.3	3.58E-02

HLA allele	Asian (%)	White (%)	*P* (Fisher)
B*35:03	8.1	2.9	3.80E-02
DQA1*01:02	27	40	3.94E-02
A*68:01	12	5.5	4.09E-02
A*23:01	0	5.1	5.90E-02
B*15:17	4.1	0.88	6.15E-02
DQB1*05:02	11	5.1	6.26E-02
B*15:01	2.7	9.3	6.80E-02
DRB3*02:10	1.4	6.9	6.86E-02
DPA1*02:01	38	27	6.95E-02
B*35:02	5.4	2	9.55E-02
A*24:07	2.7	0.44	9.70E-02
DRB1*04:04	2.7	8.8	1.01E-01
C*07:04	1.4	6.4	1.03E-01
A*01:01	23	33	1.05E-01
DQB1*03:02	12	21	1.12E-01
DRB1*04:03	4.1	1.3	1.20E-01
A*02:07	1.4	0	1.41E-01
A*26:03	1.4	0	1.41E-01
A*34:01	1.4	0	1.41E-01
A*74:01	1.4	0	1.41E-01
B*15:13	1.4	0	1.41E-01
B*27:04	1.4	0	1.41E-01
C*03:06	1.4	0	1.41E-01
DPA1*03:01	1.4	0	1.41E-01
DPA1*04:01	1.4	0	1.41E-01
DQB1*04:01	1.4	0	1.41E-01
DRB1*11:02	1.4	0	1.41E-01
DQB1*02:02	24	17	1.47E-01
DRB1*16:02	2.7	0.66	1.47E-01
DRB3*03:01	32	24	1.51E-01
C*16:01	1.4	5.5	1.54E-01
A*29:02	1.4	5.8	1.54E-01
DQA1*04:01	1.4	6	1.57E-01
DRB1*15:01	22	30	1.67E-01
B*40:01	14	8.6	1.95E-01
DRB1*08:01	1.4	4.9	2.29E-01
B*38:01	0	3.1	2.36E-01
DRB1*13:01	16	11	2.41E-01
DRB5*99:01	95	98	2.46E-01
A*03:02	1.4	0.22	2.62E-01
B*35:20	1.4	0.22	2.62E-01
B*38:02	1.4	0.22	2.62E-01
B*48:01	1.4	0.22	2.62E-01
DRB1*15:03	1.4	0.22	2.62E-01

HLA allele	Asian (%)	White (%)	*P* (Fisher)
DQB1*05:01	15	21	2.74E-01
B*08:01	16	22	2.86E-01
C*01:02	9.5	6.2	3.12E-01
B*50:01	2.7	1.3	3.13E-01
DRB3*02:02	59	53	3.15E-01
B*18:01	4.1	8	3.38E-01
B*35:08	1.4	0.44	3.66E-01
DRB1*08:03	1.4	0.44	3.66E-01
DRB1*09:01	2.7	1.5	3.68E-01
DRB4*01:01	41	47	3.79E-01
DRB1*11:01	12	8.6	3.81E-01
DQA1*02:01	30	25	3.87E-01
DRB1*07:01	30	25	3.87E-01
A*25:01	0	2.7	3.90E-01
B*58:01	4.1	2.2	4.08E-01
DRB1*13:02	4.1	6.9	4.54E-01
B*47:01	1.4	0.66	4.56E-01
DRB5*01:01	24	29	4.87E-01
DRB5*02:02	5.4	3.8	5.18E-01
C*06:02	22	18	5.22E-01
B*41:01	1.4	0.88	5.33E-01
B*53:01	1.4	0.88	5.33E-01
DRB1*08:02	1.4	0.88	5.33E-01
DRB1*11:04	5.4	4	5.33E-01
B*13:02	5.4	4.2	5.49E-01
DQB1*06:03	14	12	5.65E-01
B*37:01	6.8	5.1	5.74E-01
B*45:01	1.4	1.1	5.99E-01
B*56:01	1.4	1.1	5.99E-01
C*17:01	1.4	1.1	5.99E-01
DPA1*01:04	1.4	1.1	5.99E-01
B*39:06	0	1.5	6.01E-01
A*02:05	0	1.8	6.08E-01
A*30:02	0	1.8	6.08E-01
A*68:02	0	1.8	6.08E-01
B*14:01	0	1.8	6.08E-01
B*39:01	0	1.8	6.08E-01
B*49:01	0	1.8	6.08E-01
DRB1*04:02	0	1.8	6.08E-01
DRB1*04:07	0	1.8	6.08E-01
DRB1*13:03	0	1.8	6.08E-01
DRB1*01:02	0	2	6.21E-01
DRB1*12:01	0	2	6.21E-01
C*12:03	8.1	6.9	6.29E-01

HLA allele	Asian (%)	White (%)	*P* (Fisher)
DQB1*05:03	9.5	7.5	6.38E-01
B*35:01	9.5	8	6.48E-01
C*03:03	6.8	9.1	6.59E-01
DQA1*01:01	26	29	6.77E-01
DQA1*05:01	43	41	7.05E-01
DRB1*01:03	1.4	3.1	7.06E-01
DRB1*16:01	1.4	3.1	7.06E-01
DQB1*02:01	27	25	7.75E-01
DQB1*03:01	28	27	7.78E-01
DQB1*04:02	4.1	5.8	7.84E-01
C*07:02	28	31	7.85E-01
DRB4*99:01	92	90	8.31E-01
B*44:03	11	10	8.37E-01
C*03:04	11	12	8.48E-01
A*24:02	16	17	8.70E-01
DRB1*03:01	26	25	8.86E-01
DRB3*99:01	55	56	9.00E-01
A*69:01	0	0.88	1.00E+00
B*15:18	0	0.88	1.00E+00
DRB1*11:03	1.4	1.3	1.00E+00
A*01:02	0	0.22	1.00E+00
A*11:02	0	0.22	1.00E+00
A*24:03	0	0.22	1.00E+00
A*66:01	0	0.22	1.00E+00
A*80:01	0	0.22	1.00E+00
B*07:05	0	0.22	1.00E+00
B*15:04	0	0.22	1.00E+00
B*15:20	0	0.22	1.00E+00
B*15:30	0	0.22	1.00E+00
B*39:24	0	0.22	1.00E+00
B*44:05	0	0.22	1.00E+00
B*81:01	0	0.22	1.00E+00
C*15:04	0	0.22	1.00E+00
DQA1*01:04	0	0.22	1.00E+00
DQA1*05:05	0	0.22	1.00E+00
DQB1*03:04	0	0.22	1.00E+00
DQB1*05:04	0	0.22	1.00E+00
DRB1*08:04	0	0.22	1.00E+00
DRB1*14:03	0	0.22	1.00E+00
DRB1*14:44	0	0.22	1.00E+00
DRB1*14:01	5.4	6.6	1.00E+00
A*32:01	6.8	7.1	1.00E+00
DQB1*06:04	4.1	5.1	1.00E+00
B*40:02	1.4	2.4	1.00E+00

HLA allele	Asian (%)	White (%)	*P* (Fisher)
C*04:01	18	17	1.00E+00
DQB1*03:03	8.1	9.3	1.00E+00
A*33:01	0	1.3	1.00E+00
B*55:01	5.4	6	1.00E+00
B*57:01	5.4	6	1.00E+00
A*30:01	2.7	3.1	1.00E+00
DQB1*06:09	1.4	2.7	1.00E+00
DRB1*04:05	1.4	1.5	1.00E+00
A*30:04	0	0.66	1.00E+00
B*41:02	0	0.66	1.00E+00
A*29:01	0	0.44	1.00E+00
A*34:02	0	0.44	1.00E+00
B*15:03	0	0.44	1.00E+00
B*27:02	0	0.44	1.00E+00
B*51:08	0	0.44	1.00E+00
C*16:04	0	0.44	1.00E+00
DRB1*04:11	0	0.44	1.00E+00
DRB1*13:05	0	0.44	1.00E+00

Table S2. HLA frequencies in the two main self reported ethnic groups in the cohort. We performed Fisher's exact test to determine association between ethnicity and each HLA allele. The table is ordered by the p value of the Fisher's exact test. The p value is not corrected for multiple testing.

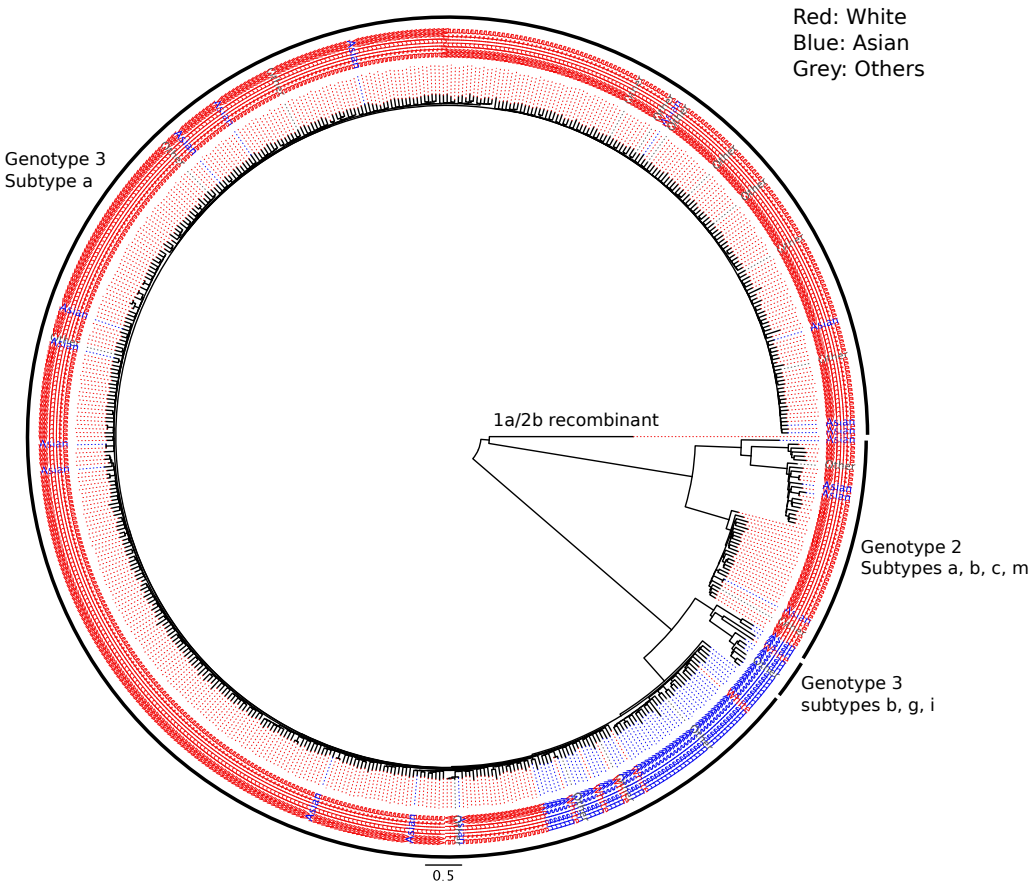


Figure S2. Virus whole-genome phylogeny labeled by self-reported ethnicity for all patients. Maximum-likelihood phylogenetic tree was estimated from virus whole-genome sequences using RAxML software with a general time reversible substitution model incorporating a gamma model of rate heterogeneity and a constrained guide tree. The resulting tree was mid-point rooted. The tips of the tree are coloured by self reported host ethnicities. White, Asian and others are coloured respectively as red, blue and grey on the tips of the tree.

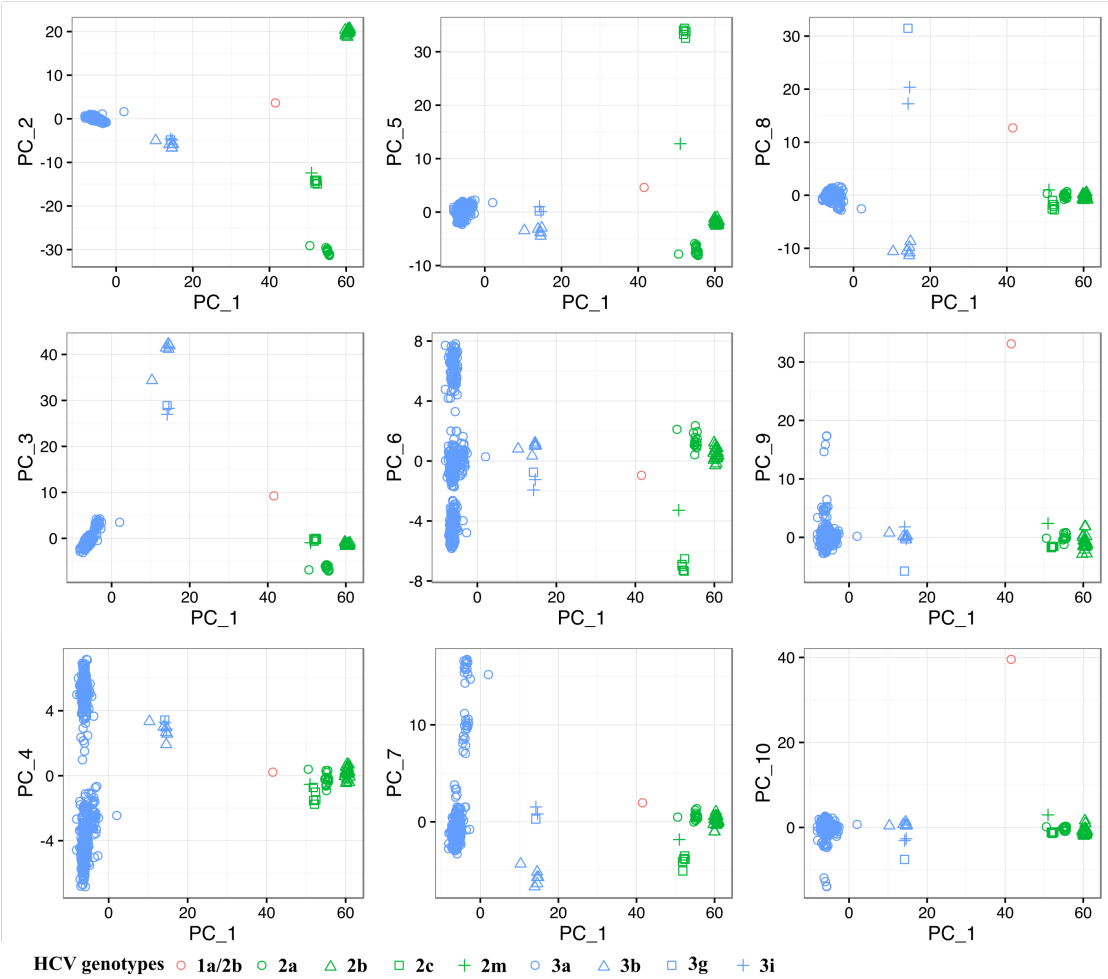


Figure S3. PCA of the virus nucleotides is associated with HCV subtypes. PCA plots (1st to 10th components) estimated from virus whole-genome sequences. HCV genotypes were inferred by comparing sequences with HCV whole genome reference sequences. Blue dots represent HCV genotype 3 (circles represent subtype 3a, triangles subtype 3b, square represents subtype 3g, crosses represent subtype 3i); green dots represent genotype 2 (circles represent subtype 2a, triangles subtype 2b, squares represent subtype 2c, cross represents subtype 2m); the red circle represents the 1a/2b recombinant isolate.

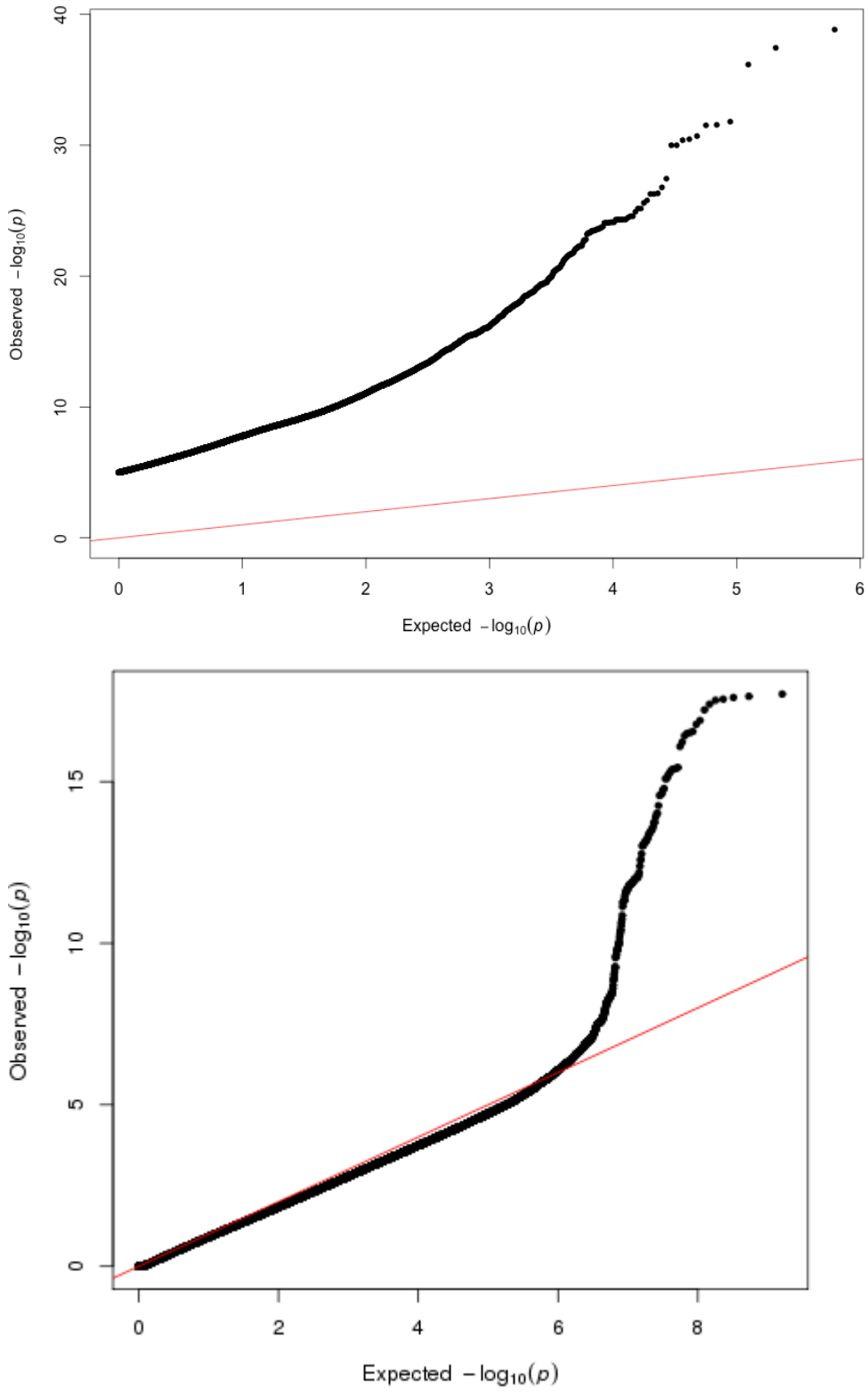


Figure S4. QQplot for genome-to-genome analysis. Top panel) No adjustment for population structure. Bottom panel) Adjusted for human (first 3 PCs) and virus (first 10 PCs) population structures

Human SNP rsID	Chromosome	Position	*P*	Associated viral amino acid variant	Viral gene
rs3129073	6	29615822	2.42E-19	F_1444	NS3
rs3117289	6	29609921	2.92E-19	F_1444	NS3
rs29228	6	29623739	3.32E-19	F_1444	NS3
rs1610742	6	29677952	6.93E-19	F_1444	NS3
rs3129039	6	29658864	7.54E-19	F_1444	NS3
rs1632962	6	29683808	8.19E-19	F_1444	NS3
rs3129063	6	29645613	9.57E-19	F_1444	NS3
rs1633063	6	29726046	1.44E-18	F_1444	NS3
rs1632964	6	29682239	1.84E-18	F_1444	NS3
rs1610601	6	29700183	2.32E-18	F_1444	NS3
rs3117295	6	29653377	3.22E-18	F_1444	NS3
rs1736922	6	29693969	3.50E-18	F_1444	NS3
rs29218	6	29607429	3.59E-18	F_1444	NS3
rs1628578	6	29695502	4.32E-18	F_1444	NS3
rs1736915	6	29704316	3.20E-17	F_1444	NS3
rs1633005	6	29764472	4.22E-17	F_1444	NS3
rs1633096	6	29707267	8.13E-17	F_1444	NS3
rs2107203	6	29669830	1.95E-16	F_1444	NS3
rs3129066	6	29668537	2.45E-16	F_1444	NS3
rs375984	6	29644502	2.50E-16	F_1444	NS3
rs1633088	6	29713580	2.57E-16	F_1444	NS3
rs2535238	6	29645038	3.10E-16	F_1444	NS3
rs3117301	6	29654700	3.15E-16	F_1444	NS3
rs2747421	6	29645118	3.77E-16	F_1444	NS3
rs3129055	6	29670261	3.87E-16	F_1444	NS3
rs1611350	6	29698821	4.11E-16	F_1444	NS3
rs2747460	6	29657127	4.37E-16	F_1444	NS3
rs356971	6	29979797	4.76E-16	F_1444	NS3
rs1633021	6	29746869	4.99E-16	F_1444	NS3
rs3116803	6	29667872	5.20E-16	F_1444	NS3
rs1264709	6	30056476	1.91E-15	F_1444	NS3
rs3129012	6	29988642	2.21E-15	F_1444	NS3
rs2523971	6	29938258	4.74E-15	F_1444	NS3
rs1611133	6	29809382	6.89E-15	F_1444	NS3
rs1611136	6	29811131	7.40E-15	F_1444	NS3
rs2844796	6	30072511	9.92E-15	F_1444	NS3
rs3131865	6	29672165	1.10E-14	F_1444	NS3

rs9257794	6	29335661	1.47E-14	F_1444	NS3
rs2508049	6	29823883	1.94E-14	F_1444	NS3
rs2517862	6	29821937	2.23E-14	F_1444	NS3
rs259940	6	30011934	2.59E-14	F_1444	NS3
rs2734970	6	29834472	2.70E-14	F_1444	NS3
rs356967	6	29973910	3.03E-14	F_1444	NS3
rs401618	6	29950210	3.23E-14	F_1444	NS3
rs2735067	6	29951106	3.82E-14	F_1444	NS3
rs6934993	6	29337500	4.01E-14	F_1444	NS3
rs356968	6	29975047	4.27E-14	F_1444	NS3
rs1150742	6	30028831	4.86E-14	F_1444	NS3
rs3094176	6	29821270	5.22E-14	F_1444	NS3
rs3129683	6	29286670	6.85E-14	F_1444	NS3
rs1245371	6	30038352	7.30E-14	F_1444	NS3
rs259943	6	30015330	7.42E-14	F_1444	NS3
rs1610682	6	29801324	9.01E-14	F_1444	NS3
rs1264697	6	30067568	1.10E-13	F_1444	NS3
rs2734986	6	29818568	2.09E-13	F_1444	NS3
rs1110446	6	30070937	2.24E-13	F_1444	NS3
rs9258593	6	29819599	2.67E-13	F_1444	NS3
rs707910	6	29910663	3.06E-13	F_1444	NS3
rs1611635	6	29836463	3.06E-13	F_1444	NS3
rs2240071	6	30070932	3.23E-13	F_1444	NS3
rs1264702	6	30065575	3.50E-13	F_1444	NS3
rs1264703	6	30065416	3.86E-13	F_1444	NS3
rs413985	6	29894600	4.05E-13	F_1444	NS3
rs1655900	6	29916618	4.14E-13	F_1444	NS3
rs2523964	6	29939294	4.83E-13	F_1444	NS3
rs1611701	6	29827956	5.66E-13	F_1444	NS3
rs1630969	6	29765001	6.81E-13	F_1444	NS3
rs2523933	6	29932292	7.28E-13	F_1444	NS3
rs2734985	6	29818662	8.64E-13	F_1444	NS3
rs2517870	6	29818744	9.01E-13	F_1444	NS3
rs1736994	6	29766596	9.28E-13	F_1444	NS3
rs1633004	6	29764663	1.01E-12	F_1444	NS3
rs2734980	6	29821896	1.07E-12	F_1444	NS3
rs5013088	6	29820389	1.07E-12	F_1444	NS3
rs259919	6	30025503	1.12E-12	F_1444	NS3
rs1737041	6	29736229	1.26E-12	F_1444	NS3
rs1611208	6	29759885	1.33E-12	F_1444	NS3
rs1633041	6	29733223	1.40E-12	F_1444	NS3

rs1077433	6	29783064	1.46E-12	F_1444	NS3
rs2734984	6	29820831	1.55E-12	F_1444	NS3
rs1736959	6	29782470	1.58E-12	F_1444	NS3
rs1736963	6	29777983	1.58E-12	F_1444	NS3
rs885940	6	29739057	1.58E-12	F_1444	NS3
rs2517861	6	29821982	1.80E-12	F_1444	NS3
rs1116221	6	30071330	1.94E-12	F_1444	NS3
rs1736983	6	29768049	2.10E-12	F_1444	NS3
rs1610637	6	29739906	2.11E-12	F_1444	NS3
rs1611209	6	29759923	2.11E-12	F_1444	NS3
rs1632988	6	29772395	2.14E-12	F_1444	NS3
rs1615962	6	29745726	2.19E-12	F_1444	NS3
rs1736971	6	29776322	2.25E-12	F_1444	NS3
rs2734945	6	29855945	2.58E-12	F_1444	NS3
rs1108052	6	30069795	2.59E-12	F_1444	NS3
rs1629068	6	29765291	2.61E-12	F_1444	NS3
rs1632987	6	29772549	2.71E-12	F_1444	NS3
rs209181	6	28792477	2.73E-12	F_1444	NS3
rs1264695	6	30067892	2.85E-12	F_1444	NS3
rs259939	6	30011581	4.69E-12	F_1444	NS3
rs1632973	6	29777171	5.51E-12	F_1444	NS3
rs1610663	6	29778626	6.13E-12	F_1444	NS3
rs2523408	6	29778109	6.21E-12	F_1444	NS3
rs2765220	6	28769671	9.11E-12	F_1444	NS3
rs1610699	6	29778987	9.95E-12	F_1444	NS3
rs1610707	6	29782048	1.42E-11	F_1444	NS3
rs259938	6	30008003	2.23E-11	F_1444	NS3
rs1130356	6	29796327	2.98E-11	F_1444	NS3
rs1737049	6	29735104	4.08E-11	F_1444	NS3
rs445150	6	29646879	5.05E-11	F_1444	NS3
rs2747430	6	29648506	5.35E-11	F_1444	NS3
rs396660	6	29646165	6.71E-11	F_1444	NS3
rs1233495	6	29444121	9.22E-11	F_1444	NS3
rs9257182	6	28754817	1.10E-10	F_1444	NS3
rs1233487	6	29468809	1.24E-10	F_1444	NS3
rs916570	6	30066031	1.24E-10	F_1444	NS3
rs2844531	6	31353171	1.53E-10	T_1646	NS3
rs1633030	6	29745794	1.58E-10	F_1444	NS3
rs2735042	6	29747373	1.58E-10	F_1444	NS3
rs1233367	6	29622220	1.62E-10	F_1444	NS3
rs1614309	6	29745884	2.49E-10	F_1444	NS3

rs2844559	6	31340075	2.49E-10	T_1646	NS3
rs3094013	6	31434366	2.56E-10	T_1646	NS3
rs1737031	6	29737481	2.85E-10	F_1444	NS3
rs2844531	6	31353171	5.22E-10	T_1646	NS3
rs3131619	6	31434331	5.38E-10	T_1646	NS3
rs11758516	6	30043486	5.50E-10	R_1270	NS3
rs2844805	6	29940679	5.60E-10	F_1444	NS3
rs1633028	6	29746367	5.93E-10	F_1444	NS3
rs3134792	6	31312326	6.40E-10	T_1646	NS3
rs9393989	6	30040084	6.98E-10	F_1873	NS4B
rs3131619	6	31434331	8.43E-10	T_1646	NS3
rs11758516	6	30043486	8.53E-10	R_1270	NS3
rs9295825	6	29941431	8.97E-10	F_1873	NS4B
rs1625792	6	31306420	9.09E-10	T_1646	NS3
rs4711207	6	30005754	9.17E-10	F_1873	NS4B
rs4947244	6	29954364	9.17E-10	F_1873	NS4B
rs9393986	6	29973526	9.17E-10	F_1873	NS4B
rs2844559	6	31340075	9.47E-10	T_1646	NS3
rs3094013	6	31434366	9.82E-10	T_1646	NS3
rs1610724	6	29761516	1.04E-09	F_1444	NS3
rs1736951	6	29785776	1.07E-09	F_1444	NS3
rs4711206	6	29987184	1.08E-09	F_1873	NS4B
rs9380150	6	30010492	1.08E-09	F_1873	NS4B
rs1611704	6	29828467	1.12E-09	F_1873	NS4B
rs9295829	6	30028800	1.14E-09	F_1873	NS4B
rs9357092	6	29984252	1.20E-09	F_1873	NS4B
rs1150735	6	30045199	1.25E-09	F_1444	NS3
rs3094605	6	31430694	1.36E-09	T_1646	NS3
rs1736946	6	29786732	1.41E-09	F_1444	NS3
rs2844531	6	31353171	1.42E-09	M_1646	NS3
rs2596565	6	31353329	1.44E-09	T_1646	NS3
rs1737004	6	29761211	1.45E-09	F_1444	NS3
rs4713274	6	29937493	1.57E-09	F_1873	NS4B
rs3134792	6	31312326	1.70E-09	T_1646	NS3
rs3094605	6	31430694	1.72E-09	T_1646	NS3
rs3094605	6	31430694	1.98E-09	M_1646	NS3
rs1611704	6	29828467	2.22E-09	L_1873	NS4B
rs3131619	6	31434331	2.38E-09	M_1646	NS3
rs2844559	6	31340075	2.46E-09	M_1646	NS3
rs1625792	6	31306420	2.59E-09	T_1646	NS3
rs3094013	6	31434366	2.64E-09	M_1646	NS3

rs2596565	6	31353329	3.26E-09	M_1646	NS3
rs2523399	6	29700079	3.28E-09	F_1444	NS3
rs1059449	6	29910698	3.63E-09	R_1270	NS3
rs76086322	6	29832724	3.67E-09	F_1873	NS4B
rs12193110	6	29937104	3.69E-09	F_1873	NS4B
rs2735054	6	29700615	3.75E-09	F_1444	NS3
rs11755961	6	29953099	3.82E-09	R_1270	NS3
rs2517930	6	29745075	3.84E-09	F_1873	NS4B
rs12979860	19	39738787	1.98E-09	V_2570	NS5B

Table S3: **Host and pathogen suggestive associations** ($P<4x10^{-9}$).

HLA allele	HCV amino acid position	Viral protein	Variable amino acids at this site	Associated amino acid	*P*	q value	In a known epitope?
A*01:01	1444	NS3	FY	F	9.63E-33	< 4.00E-04	Yes
B*08:01	1646	NS3	MTAVSI	T	2.89E-12	< 4.00E-04	No
A*68:02	1344	NS3	TVA	T	3.68E-10	< 4.00E-04	No
A*02:01	1873	NS4B	LFKICVPRMTA	F	4.06E-10	< 4.00E-04	No
A*31:01	1270	NS3	RKHS	H	4.21E-10	< 4.00E-04	Yes
A*68:02	1341	NS3	VA	V	2.43E-09	4.44E-04	No
A*68:02	1452	NS3	IV	I	6.98E-09	4.44E-04	No
A*32:01	2540	NS5B	RSKHNQC	R	7.45E-09	6.00E-04	Yes
A*32:01	2537	NS5B	NDHSYEAI	N	1.13E-08	7.27E-04	Yes
B*51:01	1380	NS3	ILV	L	3.22E-08	1.67E-03	Yes
A*31:01	2510	NS5B	AKQSEMGLTV	A	4.96E-08	2.14E-03	No
B*13:02	1635	NS3	ITVLAF	T	7.38E-08	3.13E-03	No
B*38:01	2034	NS5A	STNLPIAQVDKEMGRFW	P	7.56E-08	3.13E-03	Yes
B*15:01	1290	NS3	KPARS	R	1.25E-07	4.47E-03	Yes
B*07:02	42	C	PL	P	1.55E-07	5.00E-03	Yes
A*33:03	1282	NS3	NVTSA	T	1.73E-07	5.26E-03	Yes
C*07:02	42	C	PL	P	2.27E-07	6.10E-03	Yes
A*26:01	2605	NS5B	EAGVK	G	2.61E-07	6.64E-03	Yes
C*07:01	1646	NS3	MTAVSI	T	3.01E-07	7.39E-03	No
C*15:02	561	E2	VTLI	L	5.70E-07	1.50E-02	No
DQA1*05:01	1646	NS3	MTAVSI	T	7.24E-07	1.76E-02	No
A*02:11	2600	NS5B	QKRSAL	Q	1.01E-06	3.12E-02	Yes
B*18:01	2144	NS5A	ED	E	1.16E-06	3.28E-02	Yes
A*30:02	1585	NS3	YF	Y	1.19E-06	3.28E-02	No
A*02:01	744	E2	ATVS	T	2.25E-06	5.72E-02	No
DQB1*06:01	2385	NS5A	CLTIYMV	Y	2.62E-06	6.03E-02	No
B*51:01	2148	NS5A	MTVSCKLIANR	V	2.77E-06	6.07E-02	Yes
B*40:06	2152	NS5A	NH	N	3.76E-06	7.38E-02	Yes
B*41:02	109	C	PQSAL	S	4.29E-06	8.09E-02	No
DQB1*06:01	2687	NS5B	FLTYIVH	Y	4.84E-06	8.59E-02	No
DQB1*06:01	636	E2	FVIYLQ	L	5.10E-06	8.68E-02	Yes
A*33:03	330	E1	VALMITG	I	5.65E-06	9.20E-02	No
B*13:02	348	E1	IVLMAF	M	5.96E-06	9.44E-02	No
B*51:01	2713	NS5B	IMLV	L	6.35E-06	9.67E-02	No
B*51:01	561	E2	VTLI	L	6.68E-06	9.85E-02	No
DQB1*06:01	235	E1	TKAIVRQENGSDMHL	Q	6.87E-06	1.01E-01	No
DRB1*03:01	1646	NS3	MTAVSI	T	7.55E-06	1.03E-01	No
C*04:03	1873	NS4B	LFKICVPRMTA	I	7.62E-06	1.03E-01	No

B*07:02	48	C	AT	A	7.80E-06	1.03E-01	Yes
A*25:01	2821	NS5B	RKCIQL	R	8.31E-06	1.04E-01	No
B*57:01	524	E2	AVTNLDSMIREKH	V	8.55E-06	1.05E-01	No
DQB1*06:01	2355f	NS5A	NSDTGCAIYEV	G	9.20E-06	1.05E-01	No
C*12:03	2034	NS5A	STNLPIAQVDKEMGRFW	P	9.31E-06	1.05E-01	Yes
A*33:03	2566	NS5B	CS	C	1.00E-05	1.05E-01	No
C*12:02	1496	NS3	LPHQT	P	1.05E-05	1.05E-01	No
A*68:01	1587	NS3	TAVI	T	1.06E-05	1.05E-01	No
A*02:06	707	E2	SPA	A	1.06E-05	1.05E-01	Yes
A*33:03	2049	NS5A	PS	P	1.13E-05	1.06E-01	No
A*33:03	2103	NS5A	TVI	V	1.13E-05	1.06E-01	No
A*02:05	718	E2	FWLIVYCT	F	1.14E-05	1.06E-01	No
A*33:01	2813	NS5B	ASTV	V	1.21E-05	1.07E-01	No
DQB1*06:01	1699	NS4A	EQD	Q	1.21E-05	1.07E-01	No
DQB1*02:01	1646	NS3	MTAVSI	T	1.33E-05	1.13E-01	No
B*41:02	2221	NS5A	HAYGSTF	Y	1.38E-05	1.15E-01	No
A*31:01	636	E2	FVIYLQ	Y	1.43E-05	1.17E-01	Yes
A*02:06	2355f	NS5A	NSDTGCAIYEV	G	1.49E-05	1.19E-01	No
DQB1*05:02	475	E2	ATDEYNVKG	E	1.52E-05	1.20E-01	Yes
A*33:03	2747	NS5B	DQGCSEALNVR	C	1.59E-05	1.22E-01	No
B*41:02	2209	NS5A	PLVA	P	1.72E-05	1.26E-01	No
B*41:01	2548	NS5B	EAGD	E	1.86E-05	1.34E-01	Yes
C*04:03	210	E1	SGNDRKATQ	G	1.88E-05	1.34E-01	No
C*14:02	2148	NS5A	MTVSCKLIANR	V	1.91E-05	1.34E-01	Yes
DRB1*12:01	1507	NS3	RK	R	1.93E-05	1.34E-01	Yes
B*35:02	2457	NS5B	VI	V	2.08E-05	1.48E-01	No
DQB1*06:01	1250	NS3	NTKSDH	S	2.38E-05	1.65E-01	Yes
DQB1*06:01	1630	NS3	PAS	A	2.44E-05	1.65E-01	No
DQB1*06:01	2169	NS5A	SEAFTCDHGL	T	2.44E-05	1.65E-01	No
B*49:01	1113	NS3	AS	A	2.64E-05	1.69E-01	No
A*33:03	934	NS2	FIVLCMA	M	2.64E-05	1.69E-01	Yes
DQB1*06:01	71	C	SRTPNA	R	2.65E-05	1.69E-01	Yes
B*13:01	348	E1	IVLMAF	V	2.69E-05	1.69E-01	No
A*33:03	2369a	NS5A	VAMEDTLGQ	D	2.80E-05	1.74E-01	No
DQB1*06:01	2345	NS5A	ARQKLSVT	L	2.85E-05	1.74E-01	No
A*68:01	1087	NS3	SNTAP	S	2.93E-05	1.75E-01	No
DQB1*06:01	2355e	NS5A	EDGTAKQP	T	2.95E-05	1.75E-01	No
A*33:03	2675	NS5B	SA	S	2.98E-05	1.75E-01	No

B*13:01	454	E2	QEHDGARLSYKP	D	2.99E-05	1.75E-01	Yes
A*68:02	1177	NS3	AVT	A	3.05E-05	1.75E-01	No
B*18:01	1192	NS3	AST	A	3.14E-05	1.76E-01	Yes
B*15:01	2470	NS5B	KR	K	3.19E-05	1.77E-01	No
A*33:03	2500	NS5B	AG	A	3.31E-05	1.81E-01	Yes
B*57:01	1759	NS4B	AVQTINGSD	A	3.38E-05	1.83E-01	No
DQB1*06:01	1019	NS2	RTKIAVQ	K	3.53E-05	1.86E-01	No
C*01:02	2088	NS5A	NSDTGE	D	3.53E-05	1.86E-01	No
DQB1*06:01	2320	NS5A	GSPAKDRVTNQ	R	3.58E-05	1.86E-01	Yes
A*33:03	529	E2	WFY	W	3.62E-05	1.86E-01	No
A*02:11	1875	NS4B	TSNAG	N	3.68E-05	1.86E-01	Yes
C*14:02	1960	NS4B	LVM	L	3.69E-05	1.86E-01	No
B*40:01	2247	NS5A	EDNSKMGQ	D	4.11E-05	1.96E-01	No
B*15:18	2218	NS5A	HNQ	Q	4.11E-05	1.96E-01	No
A*33:03	2033	NS5A	AKSG	K	4.16E-05	1.96E-01	Yes
A*33:03	962	NS2	AQTSNEDGKV	E	4.20E-05	1.96E-01	No
A*33:03	763	P7	HNYQSR	Q	4.25E-05	1.96E-01	No
DQB1*06:02	190	C	ASTVD	A	4.26E-05	1.96E-01	No
B*48:01	1378	NS3	KR	K	4.26E-05	1.96E-01	Yes
B*41:02	2799	NS5B	KRNGDMHQ	K	4.28E-05	1.96E-01	No
B*07:02	47	C	RCG	R	4.42E-05	1.96E-01	Yes
DQB1*06:01	2797	NS5B	KRANQETSV	A	4.43E-05	1.96E-01	No
DQB1*06:01	2051	NS5A	TMKLARIS	L	4.48E-05	1.96E-01	No
DQB1*06:01	2682	NS5B	CVIR	I	4.48E-05	1.96E-01	No
A*33:03	872	NS2	GSA	S	4.48E-05	1.96E-01	No
DQB1*06:01	580	E2	FLIYSVKHTRMCAND	K	4.51E-05	1.96E-01	No
DQB1*06:01	2861	NS5B	RQKC	K	4.51E-05	1.96E-01	No
DQB1*06:01	332	E1	MLI	L	4.53E-05	1.96E-01	No
DQB1*06:01	16	C	INVTPSHLAY	P	4.55E-05	1.96E-01	No
A*33:03	2618	NS5B	QAKSRG	K	4.56E-05	1.96E-01	No
A*33:03	271	E1	MLVIATF	M	4.63E-05	1.96E-01	No
DQB1*06:01	49	C	TAVPS	V	4.68E-05	1.96E-01	Yes
DQA1*01:01	2486	NS5B	TSNIAKVD	T	4.71E-05	1.96E-01	Yes
B*35:01	2754	NS5B	ARTVQESI	A	4.74E-05	1.96E-01	No
B*49:01	1073	NS3	TSCPAIV	S	4.80E-05	1.96E-01	No
A*33:03	2466	NS5B	SACTPFV	A	4.81E-05	1.96E-01	No
B*14:01	2047	NS5A	ATVS	V	4.82E-05	1.96E-01	No
A*68:01	1846	NS4B	RK	R	4.86E-05	1.96E-01	Yes
A*29:02	442	E2	FILV	F	4.90E-05	1.96E-01	No
C*16:01	1986	NS5A	TSMILAGE	T	5.04E-05	1.96E-01	No

B*35:01	2220	NS5A	PKMLVNTHSRID	P	5.24E-05	1.96E-01	No
DQA1*03:01	785	P7	LMVAI	M	5.32E-05	1.96E-01	No
B*08:01	1444	NS3	FY	F	5.41E-05	1.96E-01	Yes
A*33:03	2305	NS5A	DGATN	D	5.44E-05	1.96E-01	No
A*02:05	786	P7	VAITML	V	5.51E-05	1.96E-01	No
B*08:01	2955	NS5B	NKRTH	K	5.51E-05	1.96E-01	Yes
B*18:01	2995	NS5B	CSGFY	C	5.54E-05	1.96E-01	No
DQB1*06:01	2809	NS5B	TEIAVKD	E	5.58E-05	1.96E-01	No
DQB1*06:01	2732	NS5B	DTVSIEY	S	5.58E-05	1.96E-01	No
DQB1*06:01	1140	NS3	ILV	I	5.69E-05	1.96E-01	Yes
DQB1*06:01	2374	NS5A	GETRLDSPVIAK	D	5.77E-05	1.96E-01	No
B*13:01	2366	NS5A	TASGINDVP	A	5.84E-05	1.96E-01	No
B*37:01	1078	NS3	MLTIVC	T	5.86E-05	1.96E-01	No
B*51:01	553	E2	TAVS	A	5.91E-05	1.96E-01	No
DQB1*06:01	2184	NS5A	EAQ	E	6.17E-05	1.96E-01	No
A*02:06	2136	NS5A	PTSIMARVQLK	S	6.17E-05	1.96E-01	No
DQB1*06:02	1703	NS4A	QERK	R	6.18E-05	1.96E-01	No
A*33:03	2051	NS5A	TMKLARIS	L	6.21E-05	1.96E-01	No
A*33:03	1266	NS3	SATMVI	M	6.21E-05	1.96E-01	Yes
A*33:03	1027	NS3	AS	A	6.21E-05	1.96E-01	No
A*33:03	1019	NS2	RTKIAVQ	K	6.24E-05	1.96E-01	No
A*33:03	2858	NS5B	ITVQMSNL	Q	6.24E-05	1.96E-01	No
A*33:03	2861	NS5B	RQKC	K	6.24E-05	1.96E-01	No
A*33:03	1368	NS3	GST	T	6.24E-05	1.96E-01	Yes
A*33:03	2087	NS5A	AST	S	6.24E-05	1.96E-01	No
A*33:03	2184	NS5A	EAQ	A	6.24E-05	1.96E-01	No
A*33:03	16	C	INVTPSHLAY	P	6.28E-05	1.96E-01	No
A*33:03	1221	NS3	QEHNM	E	6.31E-05	1.96E-01	No
A*33:03	2320	NS5A	GSPAKDRVTNQ	R	6.31E-05	1.96E-01	Yes
A*33:03	2430	NS5B	LI	L	6.31E-05	1.96E-01	No
A*33:03	187	C	HVICLQTYPNS	C	6.35E-05	1.96E-01	No
DRB1*08:03	385	E2	TAIS	A	6.36E-05	1.96E-01	No
A*33:03	2536	NS5B	IVM	M	6.42E-05	1.96E-01	No
A*33:03	49	C	TAVPS	V	6.42E-05	1.96E-01	Yes
DRB1*11:04	590	E2	PH	P	6.43E-05	1.96E-01	No
B*49:01	312	E1	HTARQ	H	6.56E-05	1.99E-01	No
A*11:01	574	E2	GERAKDSTVPMIHL	E	6.64E-05	2.00E-01	No

Table S4. Associations between HLA alleles and viral amino acids at a 20% false discovery rate. For each combination of HLA allele and viral site, only the

most significant associated amino acid at the site is reported. The amino acids in the "variable amino acids at this site" column are sorted by decreasing frequency such that the first shown amino acid is the most common amino acid at the site.

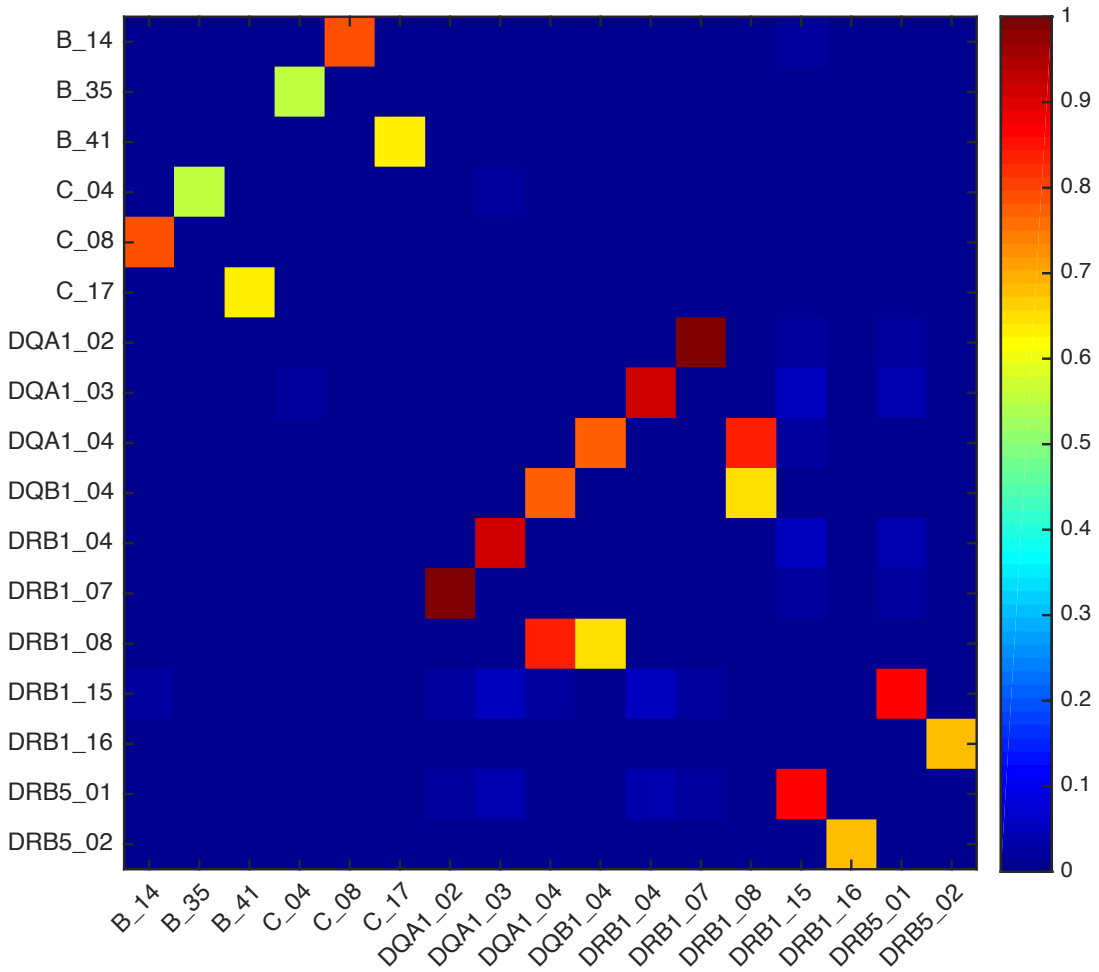


Figure S5. Linkage disequilibrium between the HLA alleles. Only alleles that present at least one pairwise $r^2>0.5$ are shown here.

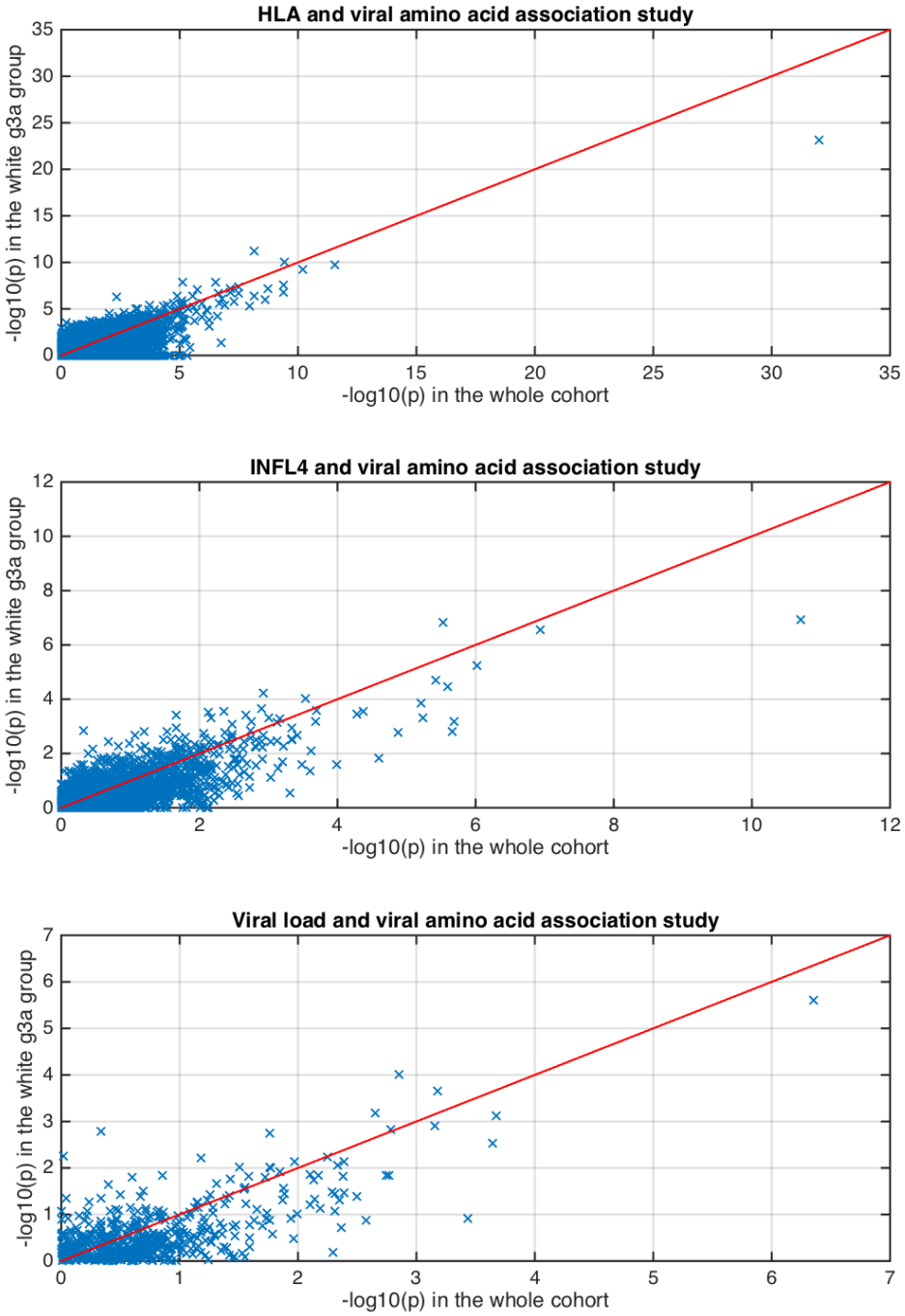


Figure S6. Scatter plot of -log$_{10}$(*P*) of association tests performed in the White ancestry patient infected with genotype 3a virus against the whole cohort. Top) Host HLA alleles and viral amino acids association study. Middle) *INFL4* genotypes (rs12979860, CC vs. non-CC) and viral amino acids association study. Bottom) Viral load and viral amino acids association study. Our phylogenetically corrected Fisher's exact test accounts for the confounding effect of population structure of the host and the virus, by estimating the within patient evolution of the virus. Performing the tests in a subset of the data where the host and viral ancestry are homogenous, results in highly correlated *P*-values for small *P*-values.

HLA allele	HCV amino acid position	Viral protein	Variable amino acids at this site	Associated amino acid	*P*	q value	In a known epitope?
C*07:02	42	C	PL	P	1.99E-07	1.33E-03	Yes
B*07:02	42	C	PL	P	8.07E-07	5.81E-03	Yes
B*41:02	109	C	PQSAL	S	9.84E-06	4.93E-02	No
B*13:02	348	E1	ILVM	I	1.25E-06	8.78E-03	No
C*08:02	372	E1	ATVIFC	I	9.41E-06	4.93E-02	No
A*29:02	444	E2	YHFRVWL	Y	4.02E-06	2.39E-02	No
DRB1*14:04	905	NS2	ATSC	S	5.58E-07	4.20E-03	No
A*31:01	1270	NS3	RKHS	R	6.20E-08	3.64E-04	Yes
B*15:01	1290	NS3	KRSA	R	3.64E-06	2.39E-02	Yes
A*68:02	1341	NS3	VA	V	9.93E-07	6.91E-03	No
A*68:02	1344	NS3	TVA	T	9.22E-11	<2.50E-04	No
B*51:01	1380	NS3	ILV	L	2.10E-07	1.38E-03	Yes
A*01:01	1444	NS3	FY	F	7.18E-24	<2.50E-04	Yes
A*68:02	1452	NS3	IV	I	3.84E-07	2.89E-03	No
B*13:02	1635	NS3	ITVL	T	1.65E-07	1.00E-03	No
B*08:01	1646	NS3	MTSAVI	T	2.01E-10	<2.50E-04	No
C*07:01	1646	NS3	MTSAVI	T	1.40E-08	<2.50E-04	No
B*57:01	1759	NS4B	AVTGNSDI	A	4.01E-06	2.39E-02	No
A*02:01	1873	NS4B	LFCIPRM	F	2.58E-08	2.50E-04	No
B*38:01	2034	NS5A	STLPQKVAIR MFW	P	6.35E-08	3.64E-04	Yes
C*12:03	2034	NS5A	STLPQKVAIR MFW	P	3.99E-06	2.39E-02	Yes
B*18:01	2144	NS5A	ED	E	2.36E-06	1.67E-02	Yes
B*51:01	2148	NS5A	MTVSLKIA	V	4.60E-07	3.16E-03	Yes
C*14:02	2148	NS5A	MTVSLKIA	V	2.32E-06	1.67E-02	Yes
B*40:01	2248	NS5A	TSKAN	T	9.62E-06	4.93E-02	No
DQA1*01:01	2486	NS5B	TIANSV	T	5.12E-06	2.90E-02	Yes
A*31:01	2510	NS5B	AKEQGMSV T	A	1.79E-06	1.35E-02	No
A*32:01	2537	NS5B	NDYEI	N	4.77E-06	2.75E-02	Yes
A*32:01	2540	NS5B	RHSNC	R	5.96E-12	<2.50E-04	Yes
B*51:01	2713	NS5B	ILMV	I	1.25E-08	<2.50E-04	No
A*25:01	2821	NS5B	RCQL	R	2.51E-06	1.73E-02	No

Table S5. Associations between HLA alleles and viral amino acids in patients with self reported White ancestry and infected with genotype 3a virus at a 5% false discovery rate. For each combination of HLA allele and viral site, only the most significant associated amino acid is reported. The first amino acid at "variable amino acids at this site" column indicates the most common amino acid at the site in the data.

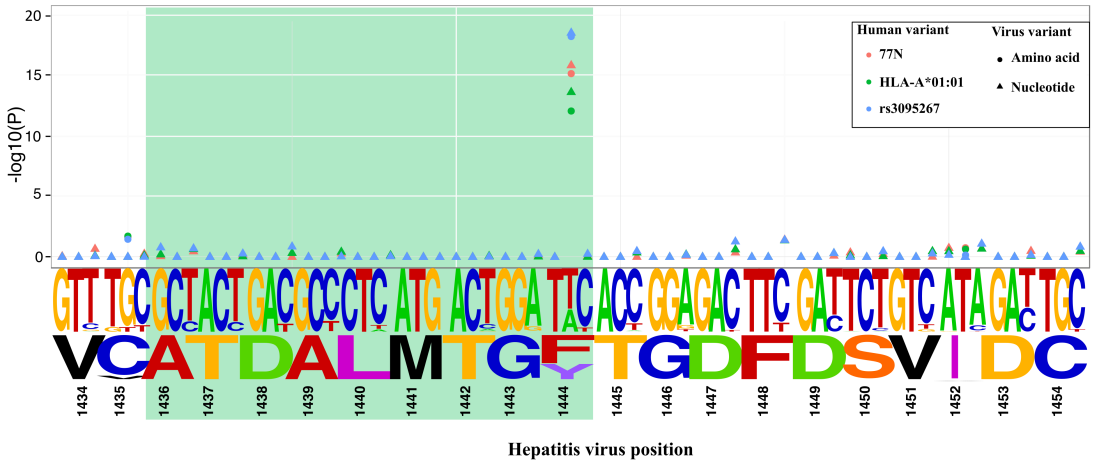


Figure S7. Sequence logo representations of the virus nucleotide and the amino acid profiles around position 1444. The dots represent the association between the human variants (SNP rs3095267 (blue), *HLA-A*01:01* (green) and amino acid N77del in HLA-A protein (red)) and the viral variants (nucleotides (triangle) and amino acids (circle)). The green region indicates the experimentally validated epitope for *HLA-A*01* allele. SNPs and HLA amino acids in the MHC region were imputed using SNP2HLA program. While rs3129073 is the most associated genotyped SNP with Y1444F; rs3095267 is the most associated imputed SNP.

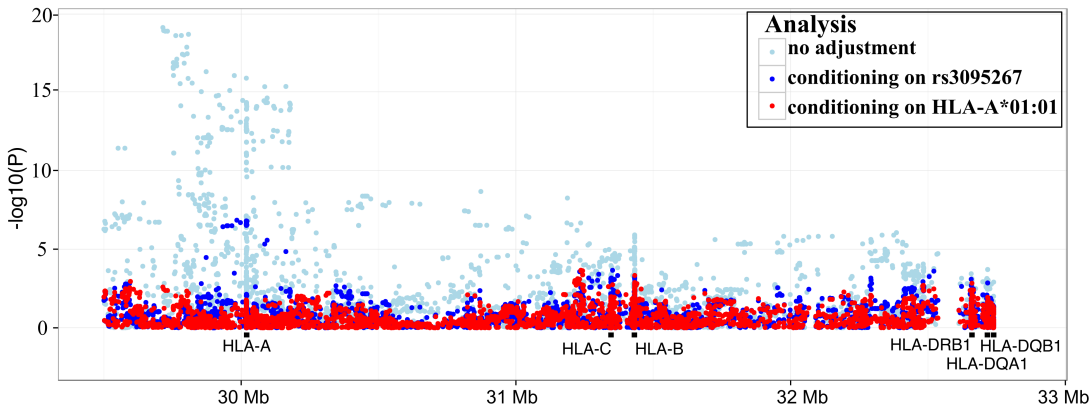


Figure S8. Association analysis across the MHC region for presence and absence of F amino acid at site 1444, whilst conditioning (including as a covariate) on genotypes at rs3095267 (top imputed SNP) and HLA-A*01:01.

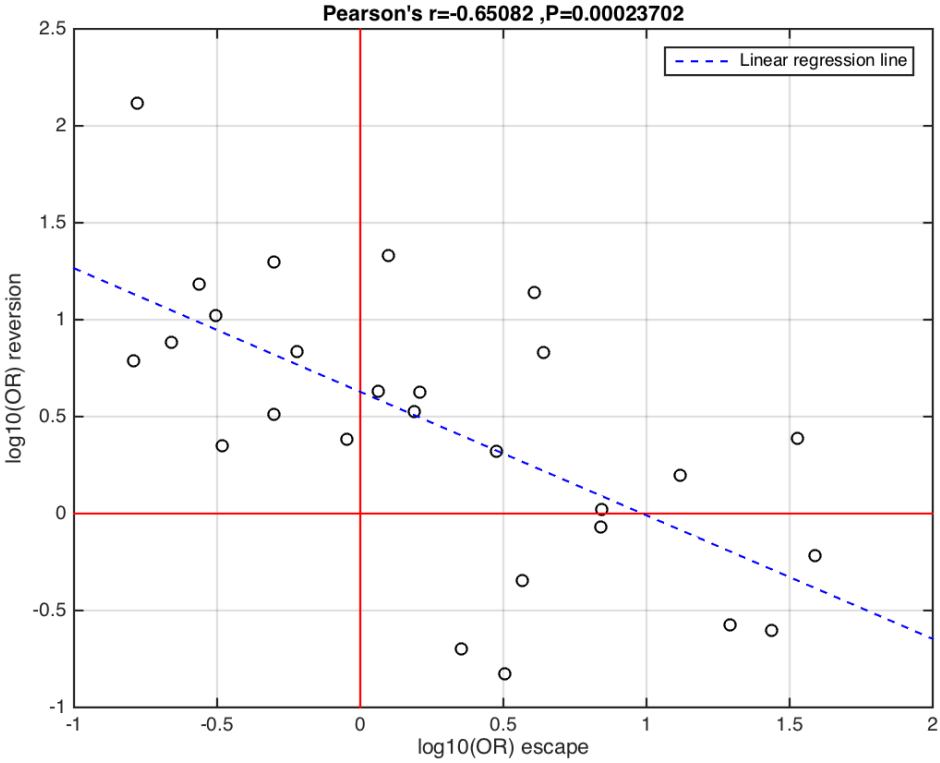


Figure S9. Bi-directionality of escape and reversion for viral amino acid sites associated with HLA alleles at 20% FDR. Here we plotted the $\log_{10}$ of odds ratio of the first part of the 2x4 contingency table against the $\log_{10}$ of odds ratio of the second part of the table only for sites that had sufficient observations to estimate the odds ratio of escape and reversion. An escape mutation away from the wild type amino acid in the people with the matched HLA alleles is usually associated with fitness cost for the virus. Therefore it is expected that in people without that HLA allele for the escape mutation to reverse to wild type. This is the assumption of bi-directionality. There is a negative correlation (R=-0.65, P<0.01) in the $\log_{10}$ of odds ratio of escape and reversion.

HCV amino acid position	Viral protein	Variable amino acids at this site	Associated amino acid	*P*	q value	In a known epitope?
2570	NS5B	AVTISGLNE	V	1.94E-11	<2.86E-04	No
2991	NS5B	HYLFVMIRW	H	1.17E-07	<2.86E-04	No
60	C	GEDRSA	G	9.63E-07	2.86E-04	No
2576	NS5B	PAST	P	2.05E-06	2.86E-04	No
576d	E2	EGDASTQNKPRVWL	E	2.16E-06	2.86E-04	No
741	E2	IVL	I	2.52E-06	2.86E-04	No
500	E2	SLARKTQPVGEI	S	5.80E-06	1.11E-03	No
2414	NS5A	SNQGEKDTHR	S	6.11E-06	1.20E-03	No
501	E2	SNTDGKERQLH	S	2.50E-05	6.77E-03	No
109	C	PQSLA	P	5.20E-05	1.35E-02	No
578	E2	DHLNSYTEGRFAICVPQK	D	1.04E-04	2.70E-02	No

Table S6. Host *IFNL4* SNP association with HCV amino acids at FDR of 5%.

Region	Beta (effect size)	95% confidence interval (beta)	*P*
HVR1	0.029	-0.024 – 0.083	2.82E-01
HVR2	0.394	0.230 – 0.559	2.53E-06
PePHD	0.155	-0.284 – 0.593	4.88E-01
ISDR	0.349	0.157 – 0.541	3.51E-04
PKR-BD	0.158	0.034 – 0.282	1.23E-02
HVR3	0.064	-0.050 – 0.179	2.71E-01

Table S7. Number of changes from the viral population consensus in *IFNL4* CC and non-CC individuals in regions of interest in genotype 3a isolates. Beta column shows the effect size on the number of changes from the consensus for patients with CC genotype relative to patients with non-CC genotype. Poisson regression was used with the counts of changes relative to consensus as the independent variable and the *IFNL4* genotypes as predictors.

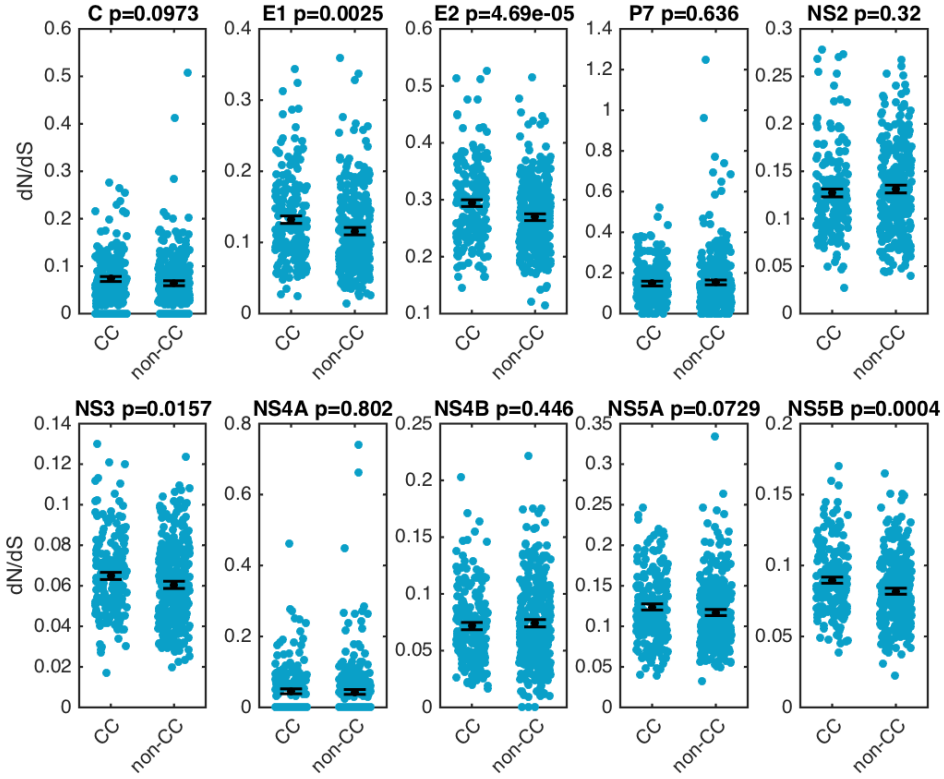


Figure S10. Association between dN/dS and *IFNL4* genotypes at the viral gene level. For each viral gene the rate of synonymous and non-synonymous mutations were estimated for all isolates. The difference in the mean dN/dS in CC and non-CC patients is nominally significant in E1, E2 NS3 and NS5B genes.

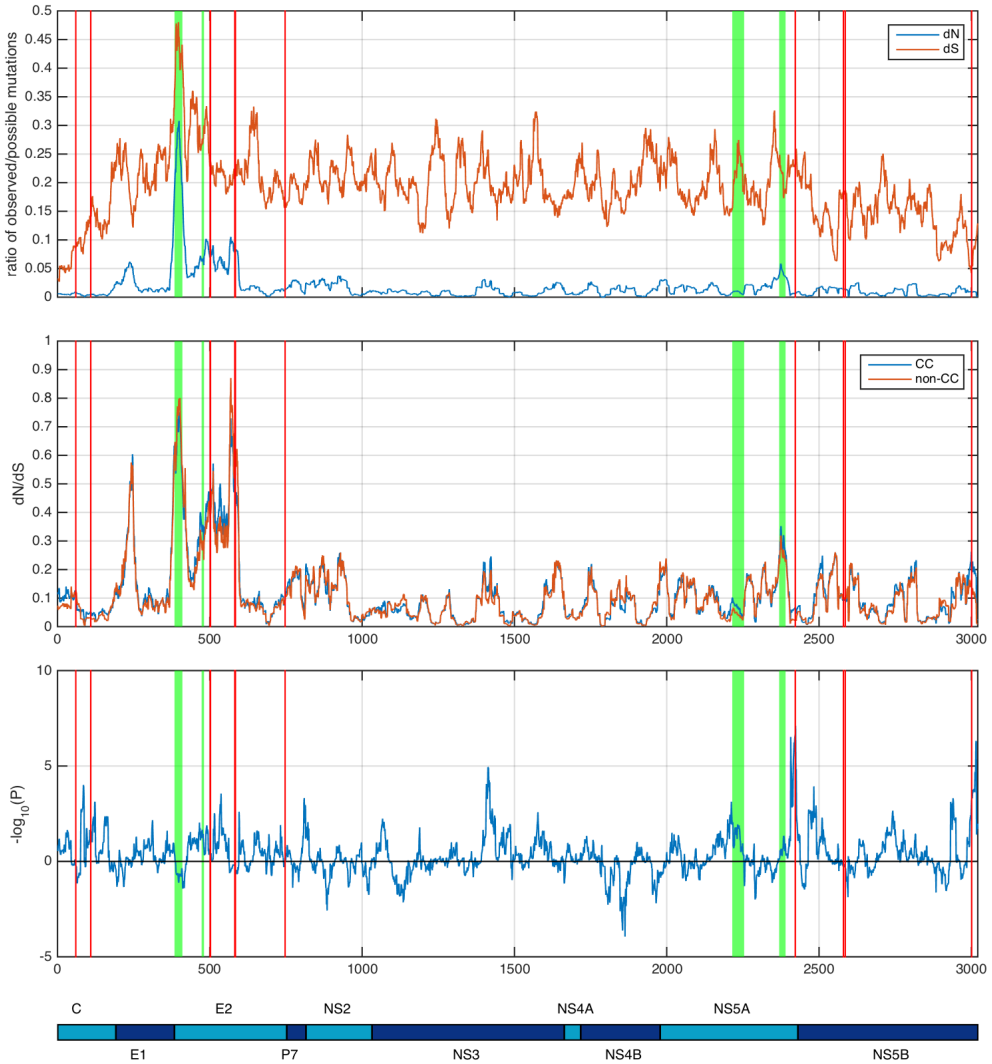


Figure S11. Sliding window analysis of dN and dS along viral genome. Using a window size of 31 nucleotides, moving one codon at a time, the pairwise rate of synonymous and non-synonymous mutations were estimated. The green regions from left to right indicate HVR1, HVR2 ISDR and HVR3 regions. The red lines indicate the 11 sites associated with *IFNL4* genotypes. Top panel) Mean rates of synonymous and non-synonymous mutations for all genotype 3a isolates. Middle panel) Mean dN/dS ratio for *IFNL4* CC and non-CC individuals. Bottom panel) Difference in mean of the dN/dS for *IFNL4* CC and non-CC individuals. The difference in mean is higher in CC individuals above 0 ($\log_{10}(P)$ is plotted) and is higher in non-CC individuals below 0 ($-\log_{10}(P)$ is plotted).

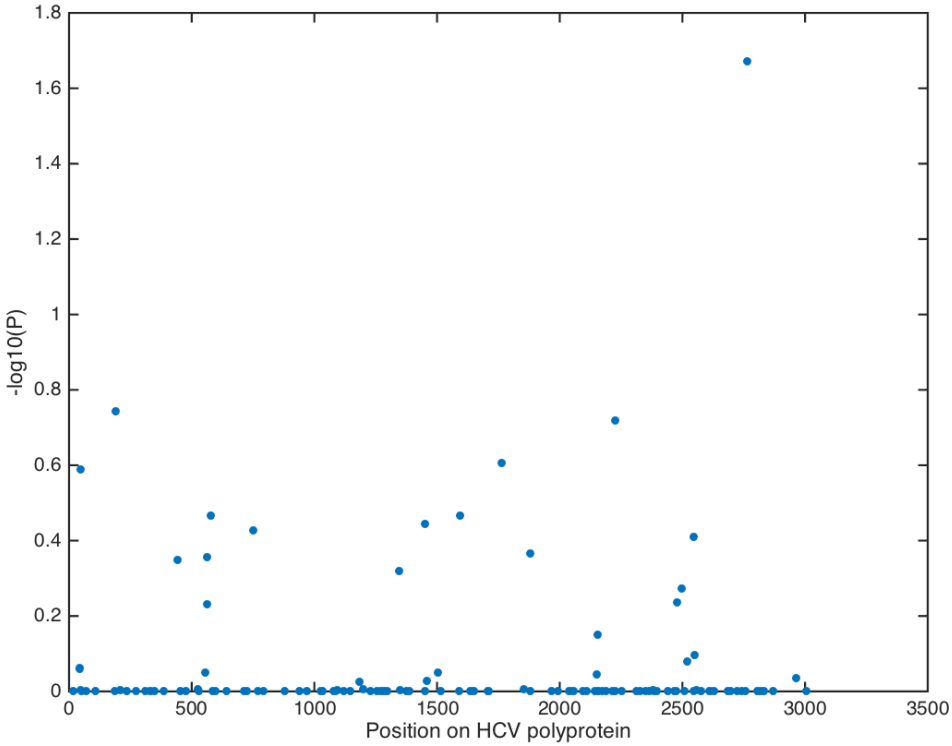
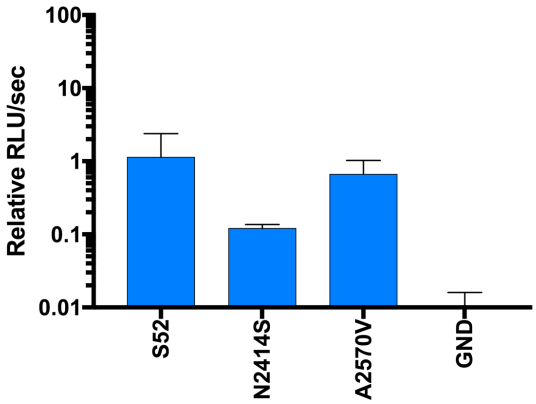
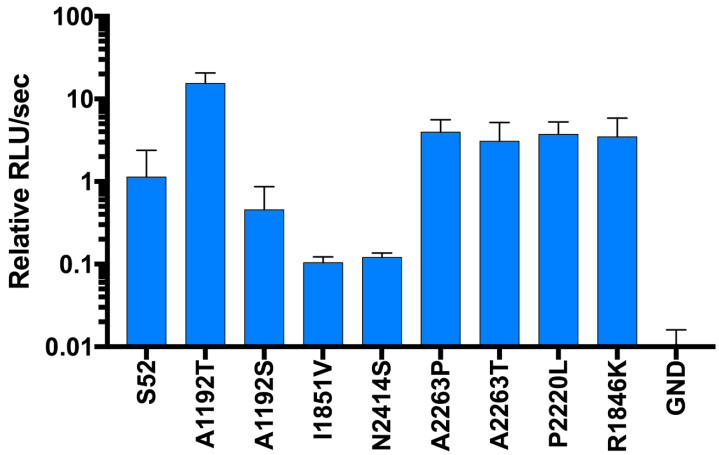


Figure S12. Interaction between *IFNL4* genotypes and HLA allele escape mutations. For the sites associated with HLA alleles at a 20% FDR, we tested for interaction between *IFNL4* genotypes and HLA allele using a logistic regression. Each dot is the $-\log_{10}(P)$ for the interaction term between *INFL4* genotypes and the HLA allele associated with the amino acids at the site.





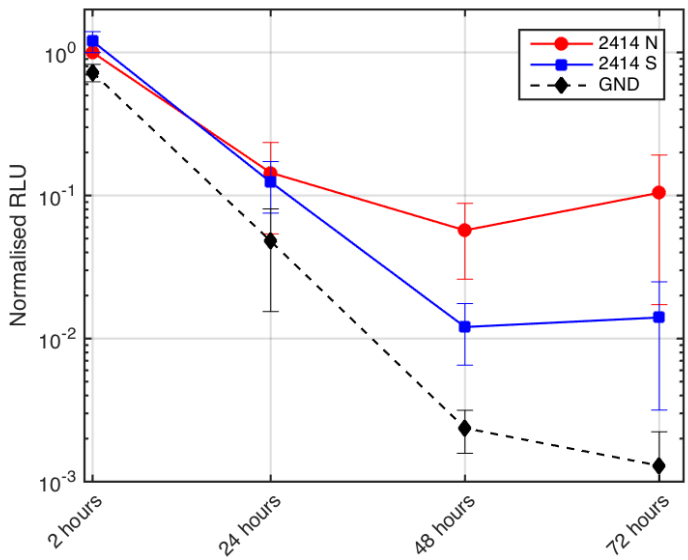


Figure S13. Genotype 3a Replicon assay for a subset of sites associated with *IFNL4* and nominally associated with viral load. Top panel, two sites associated with *IFNL4* genotypes. Middle panel, eight sites nominally associated with viral load. Bottom panel, replicon assay for site 2414 which is associated both with *IFNL4* genotypes and viral load. A change from a serine (S) to an asparagine (N) resulted in a 10-fold increase in viral replication. GND indicates negative control. Y-axis is the luciferase assay in normalized raw light units (RLU).

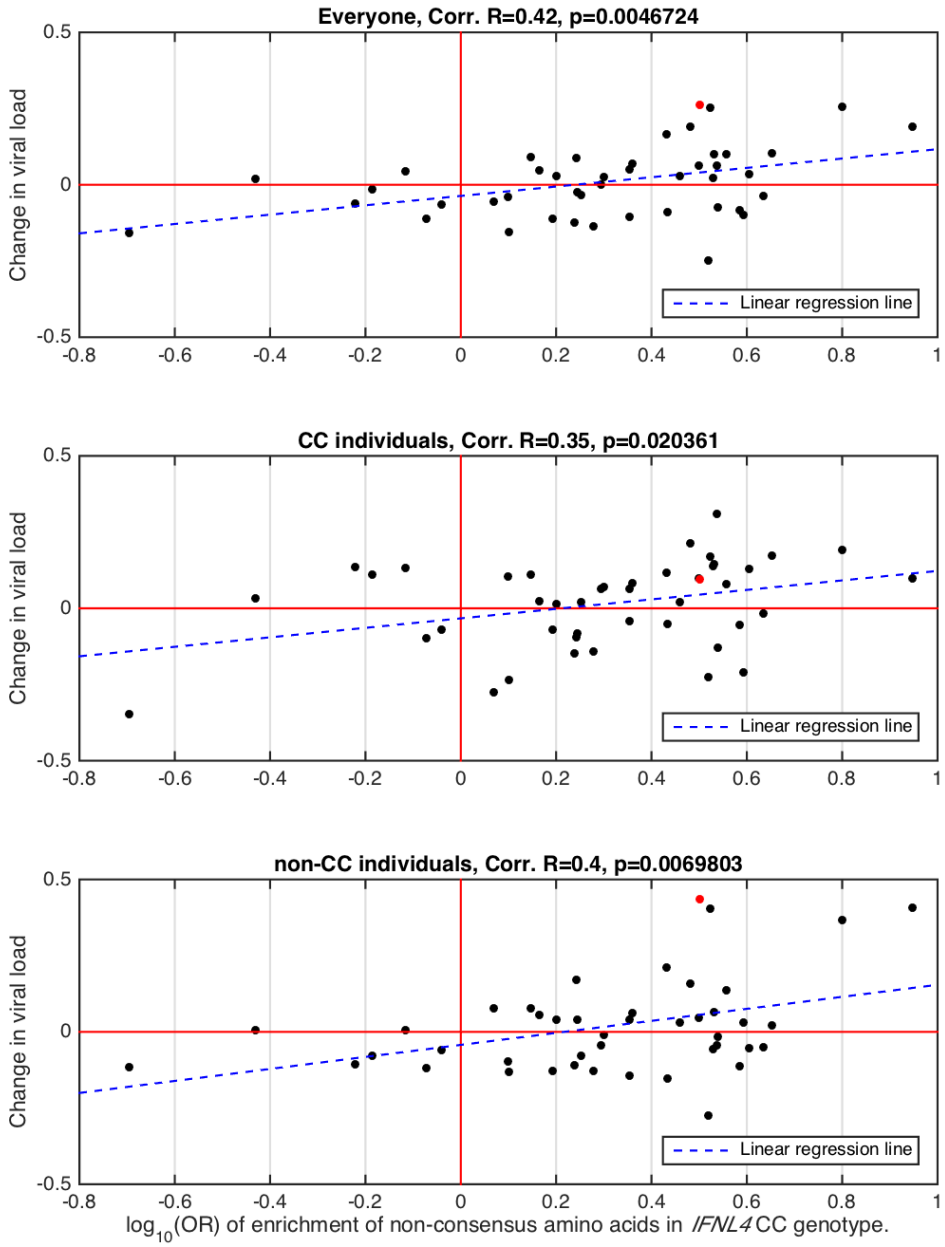


Figure S14. Viral amino acids affect on viral load vs. *IFNL4* odds ratio of escape from consensus for viral sites associated with *IFNL4* genotypes at a 20% FDR. We only plotted sites that had at least 10 non-consensus amino acids in each of *IFNL4* CC and non-CC genotypes. On x-axis we have plotted the $\log_{10}$ of odds ratio for the contingency table where rows are the counts of CC and non-CC patients and the columns are the counts for the change and no change on the terminal branches of the tree from the most common amino acid ($\log_{10}(OR) > 0$ is interpreted as proportion of changes from "most common amino acid" to "any other amino acid" is higher in CC patients than non-CC patients). On y-axis the effect size of non-consensus amino acids (relative to consensus amino acid) on viral load is plotted (estimated using linear regression where in the top panel the *IFNL4* genotypes are added as a covariate). In the bottom two panels, the effect size of the non-consensus amino acid on viral load was estimated only in *IFNL4* CC and non-CC patients. The red dot represents the viral position 2414, associated with *IFNL4* genotypes and viral load.

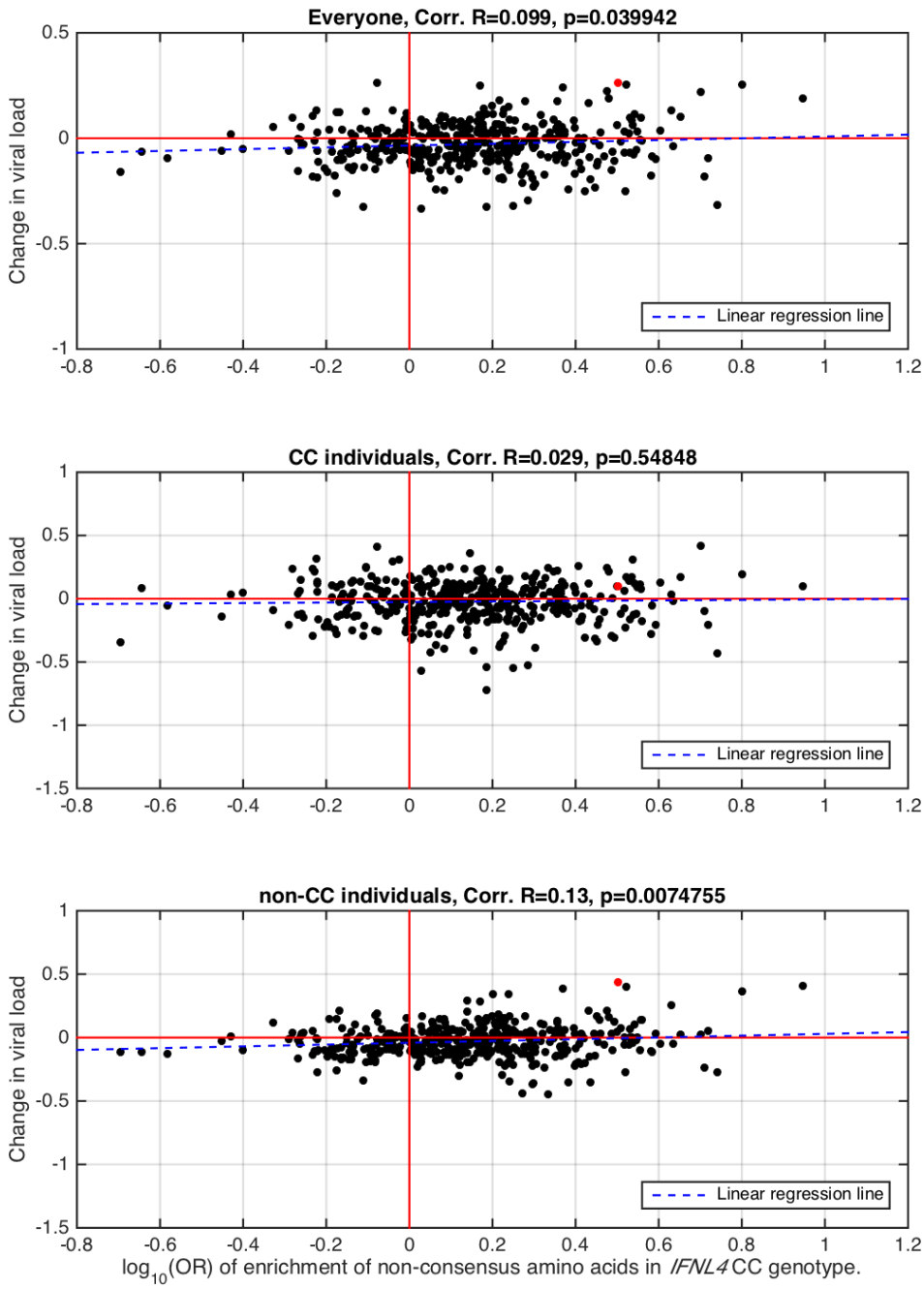


Figure S15. Viral amino acids affect on viral load vs. *IFNL4* odds ratio of escape from consensus for viral sites associated with *IFNL4* genotypes for all sites. We only plotted sites that had at least 10 non-consensus amino acids in each of *IFNL4* CC and non-CC genotypes. On x-axis we have plotted the $\log_{10}$ of odds ratio for the contingency table where rows are the counts of CC and non-CC patients and the columns are the counts for the change and no change on the terminal branches of the tree from the most common amino acid ($\log_{10}(OR) > 0$ is interpreted as proportion of changes from "most common amino acid" to "any other amino acid" is higher in CC patients than non-CC patients). On y-axis the

effect size of non-consensus amino acids (relative to consensus amino acid) on viral load is plotted (estimated using linear regression where in the top panel the *IFNL4* genotypes are added as a covariate). In the bottom two panels, the effect size of the non-consensus amino acid on viral load was estimated only in *IFNL4* CC and non-CC patients. The red dot represents the viral position 2414, associated with *IFNL4* genotypes and viral load.